BB1C – Open Orthophosphate Model – Report

V2.4 – December 2025

Table of Contents

BB1C: Open Orthophosphate Model

Transforming Water Quality Monitoring: AI-Driven Prediction of Orthophosphate Concentrations in River Catchments

Abstract

This study introduces an AI/ML-based framework for predicting orthophosphate (OrthoP) concentrations across UK water catchments. Utilising over 70 million historical observations from the Environment Agency, the model integrates molecular embeddings, correlations, spatial-temporal features, and domain expertise. The hybrid approach leverages machine learning, chemical clustering, and determinands curated by domain experts to deliver actionable water quality insights. This report demonstrates robust performance metrics (MSE=0.00088, MAE=0.02) and explainability via our SHAP[1] (Responsible AI) analysis. The performance of the Open Orthophosphate Model in the validation process (see section 16) suggests that more can still be done and explored to fully understand, and potentially improve, the performance of the model.

1. Introduction

Phosphate is an indicator of nutrient pollution in aquatic systems and is often measured as orthophosphate by regulators[2]. Measures of phosphate in freshwater bodies are essential for guiding effective nutrient management. Traditional methods of monitoring phosphate are labour intensive. As a consequence, measures of phosphate are sparse across space and time, representing less than 2% of all records in the Environment Agency (EA) database, equal to roughly 2 monthly measurements per location over 24 years. For perspective, ammonia-N (determinand notation 0111) measures represent 4.4% of all EA water quality measures in 2023.

Finding a more cost-effective solution to measuring or estimating orthophosphate, and other nutrients in water bodies, has been a key challenge in the water sector for many years that has proven difficult to improve. Any success – big or small – will be valuable in helping improve the water sectors' capabilities of understanding and reducing pollution.

To address this, River Deep Mountain AI has developed a predictive framework using Artificial Intelligence and Machine Learning (AI/ML) to estimate orthophosphate.

1 SHapley Additive exPlanations, is a Responsible AI framework, an approach to explain the output of machine learning model

2 EA WFD Cycle 2 River Phosphorous Classification, EA Focus on Phosphorous, EA & Natural England: State of the water environment

By understanding relationships between the water quality determinands in the EA database and measured orthophosphate concentration, the model can estimate orthophosphate for water quality samples in which orthophosphate was not measured. The model outputs can provide estimates of orthophosphate at a more granular spatiotemporal resolution, which can complement the sparsely monitored orthophosphate data, both historically and for measurements coming in the future, such as section 82 required water quality monitoring. This framework, leveraging 24 years of EA water quality data, can help address environmental, regulatory, and technical gaps in water quality monitoring.

Of the 70 million water quality monitoring samples from over 64,000 sampling points recorded for England in the EA water quality monitoring database over 24 years, phosphate (P) and orthophosphate account for 0.2% and 2% of those samples, respectively. This is creating gaps in real-time water quality management. To address this, the Open Orthophosphate Model leverages 70 million observations from 55 rivers and 25 aquifers[3] to develop an AI/ML framework that predicts orthophosphate concentrations in Rivers. This framework considers molecular chemistry and spatiotemporal[4] analysis to support scalable, cost-effective monitoring and regulatory compliance.

2. Data Ingestion

The project harnesses the EA water quality database, containing 24 years of water quality monitoring data, encompassing over 70 million observations. This includes 3,261 physio-chemical determinands[5], ranging from pH to nutrient concentrations, offering a holistic view of catchment water quality dynamics. Training the AI/ML models on these data offers the opportunity to decode the relationships between a wide range of water quality determinands and the concentration of orthophosphate.

3 As per EA's scope of water quality monitoring

4 Exploiting spatio patterns such as sampling point specific longitude, latitude or British National Grid, water temperature, compliance, sampling purpose and temporal patterns such as, day, year, month, week of the year, day of the week, cyclical encoding etc.,

5 determinands, determinands, variables, and features all refer to an input feature for an ML model in this document

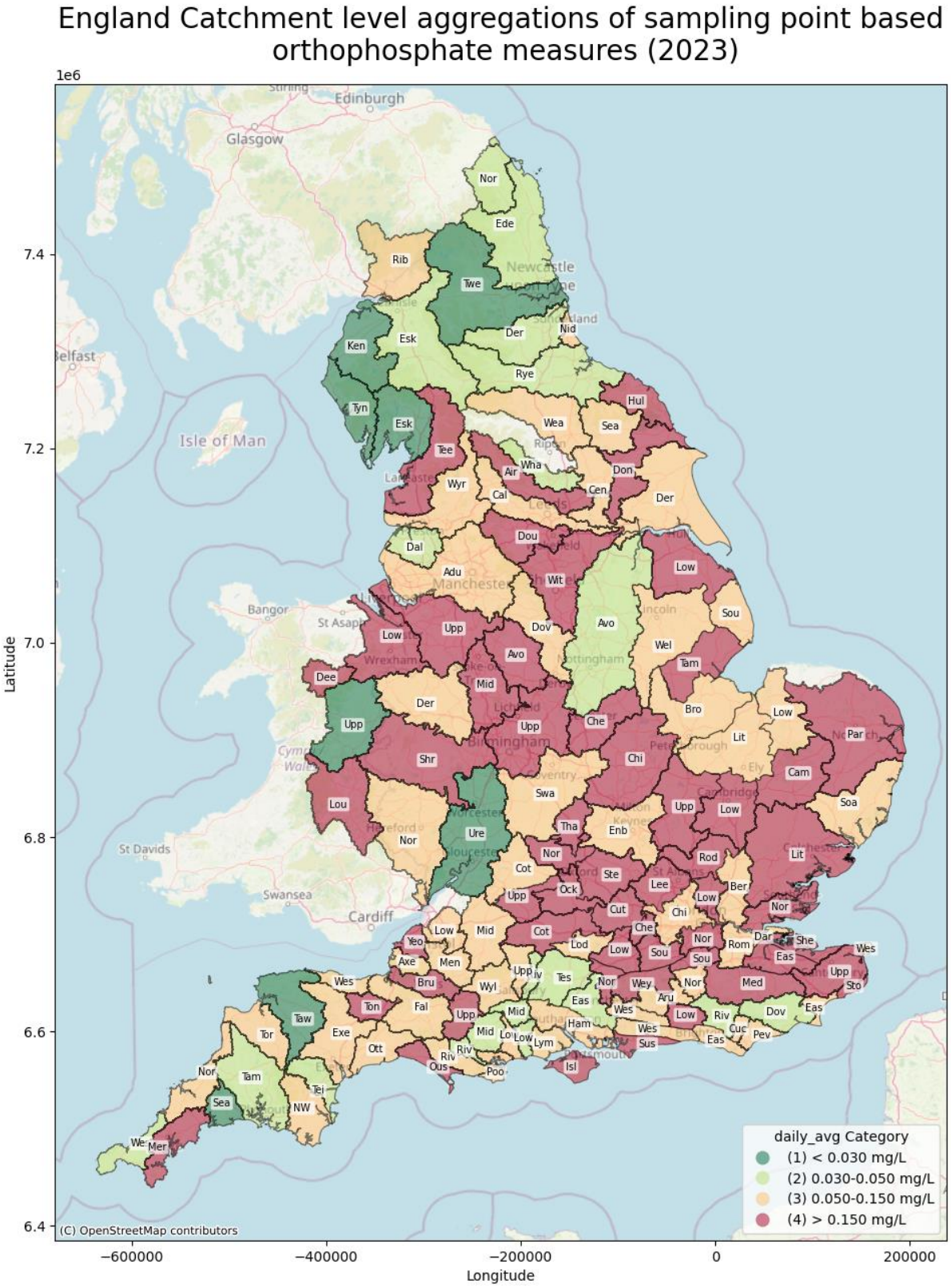


Figure 1: Map of orthophosphate concentrations across England, showing values for 2023, averaged first to daily for each monitoring point, and then averaged to a value for each hydrological area (EA Hydrological Boundaries). The classification shown broadly reflects the various boundaries for the UKTAG standards for phosphorus.

The classification within the catchments[6] in Figure 1 is based on the different phosphorus boundaries from the UKTAG standards[7] regarding phosphorus, but does not account for the variation with altitude or alkalinity.

By training the Open Orthophosphate Model on a subset of these data, the objective is for the model to be able to:

1. Predict orthophosphate levels at the 72%[8] of monitoring locations in the EA water quality database in which orthophosphate is rarely monitored or absent.
2. Improve the value of traditional water quality monitoring practices, by inferring orthophosphate levels using easily obtainable water quality determinands.

Orthophosphate distribution: Figure 2 below shows the distribution of orthophosphate present in the raw EA water quality monitoring records over the full 24-year period, including only those classified as "RIVER/SURFACE WATER". The orthophosphate concentrations range from 0.0001 mg/l to 5,000 mg/l. Here, frequency is calculated as sample counts per concentration range for a specific year. A key observation to highlight is the limit of detection (LOD) values, notably <0.02 which make up 9% of the total samples. These 'less than' values introduce an artificial stratification for the data in the lower ranges. During this project we have decided to include these samples at face value, which might influence the predictive capabilities of our final model. We have added some thoughts on this in section 13.

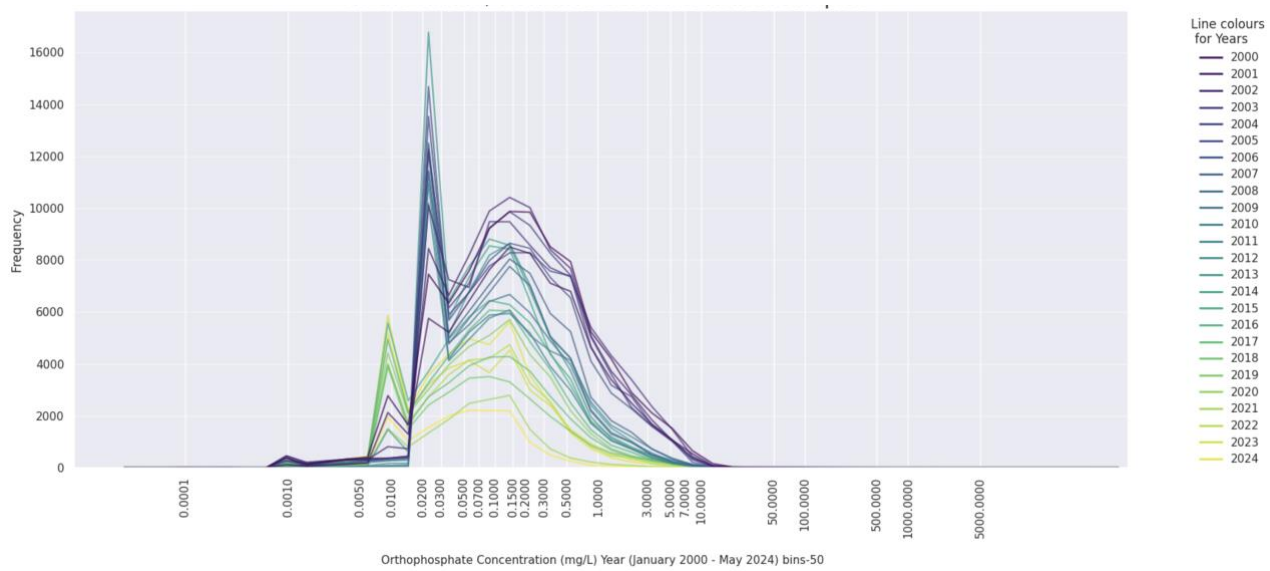


Figure 2: Distribution of orthophosphate in RIVER/SURFACE WATER samples.

6 This is as per EA hydrological boundary Click here for EA's shape file.

7 Access the updated phosphorus standards for rivers from the UK Technical Advisory Group (UKTAG) by clicking **here**

8 Orthophosphate is monitored at 18,000 sampling points out of 64,000. Using our open orthophosphate model, the orthophosphate concentration in the unmonitored locations can also be predicted.

Observations:

Concentration hotspot: Peaks between 0.01 – 0.05 mg/L suggest typical baseline levels, whereas values exceeding 0.5 mg/l may indicate pollution events.

Yearly variations: Overlapping lines for all years (2000-2024) demonstrate consistent patterns in the measurement clusters.

84.6% of the datapoints falls within the 0.5 mg/l, while 15.4% exceed it.

Together with domain experts, 0.5 mg/L was identified as the outlier cap. The next step was to assess the yearly patterns of values within this cap. The dataset assessed covered measurements from January 2000 to May 2024.

Initially, we selected a cap of 0.5 mg/L for model development, considering the measured data from various sample types. Subsequently, we decided to apply the models exclusively to surface samples from sampling points across England. However, after consulting with domain experts regarding the samples to be included, we have concluded that a revision may be necessary (see section 13).

3. Overall technical design overview

Below is a structured breakdown of the technical design and methodology of the model.

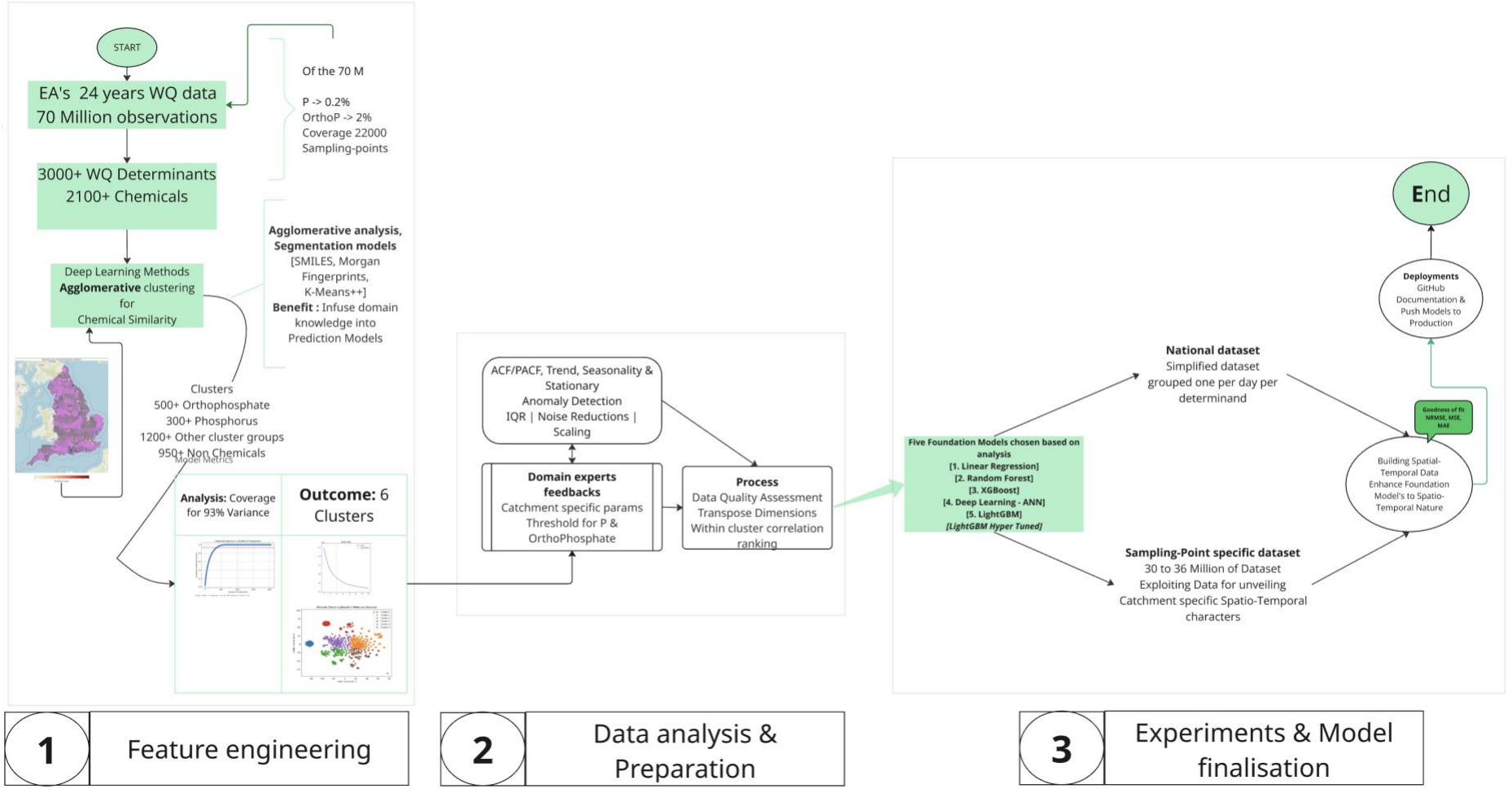


Diagram 1: Technical flow for identifying influential variables and predicting orthophosphate

3.1. Data Collection & Preprocessing (Diagram Node 1)

The process begins with 24 years of Environment Agency (EA) water quality (WQ) data, comprising 70 million observations across 2,295 chemical determinands. This phase includes data quality assessment (e.g., outlier detection via IQR[9]) methods and thresholds provided by domain experts.

Feature Engineering & Analysis: This analysis was conducted using advanced methods such as agglomerative analysis with Morgan Fingerprints for molecular similarity, K-means++ model for clustering and we have collaborated with domain experts to identify additional features relevant for orthophosphate estimation. These result in the final selection of 45 features.

3.2. Model Development (Diagram Node 2)

The 24 years water quality data was cleansed and transposed from one-dimensional format to two-dimensional model consumable format. Outliers were removed. Trend and seasonality were analysed, and additional temporal features were created accordingly. This data preparation was repeated for four types of data-cuts: one with all types of samples, for few specific sample types (such as River / Running surface water, Estuarine water, Sea water, Canal water, Pond / Lake / Reservoir water), Surface water type and Surface water without Reactive monitoring samples.

3.3. Validation & Deployment (Diagram Node 3)

We tested five different machine learning models and chose the best performing model for additional fine-tuning (see section 8). We tested a statistical model, deep learning frameworks (e.g., Artificial Neural Networks), tree-based algorithms and two gradient boosting models. Out of these, Light GBM (Gradient Boosting Model), a gradient boosting model that consistently gave low residual errors, was chosen.

The final production model has undergone spatiotemporal validation (e.g., from 8,500 training observations to 31.7 million observation exploiting sampling-point specific spatial and temporal patterns across England catchments) and has been cross-assessed against performance evaluation parameters such as NRMSE (Normalised Root Mean Square Error), MSE (Mean Square Error), MAE (Mean Absolute Error) and SHAP (SHapley Additive exPlanations) Explainable-AI.

9 The Interquartile Range (IQR) method is a statistical technique that helps identify and possibly remove unusual data points from a dataset.

4. Exploratory Data Analysis

The activities under this topic are part of the feature engineering (refer Node 1 in Diagram 1) phase. The exploratory data analysis (EDA) was conducted using two broad approaches:

1. Clustering methodology, which focused on the molecular similarity of determinands and grouping them into clusters using advanced machine learning tools. We trialled two approaches for this: ChemBERTa and agglomerative clustering, with the latter carried forward.
2. A traditional correlation approach, which analysed data points to identify significant correlations using methods such as Spearman ranking and LightGBM feature selection.

Ultimately, a hybrid approach was adopted to combine insights from both methodologies and to finalise the most representative features for the Open Orthophosphate Model.

Data Mastery & Chemical Insights: The water quality database comprises over 3,261 determinands. After excluding determinands with fewer than 100 observations in the past 24 years, 2,483 determinands were identified for further analysis. Among these, we identified 2,295 chemical determinands (i.e. two-thirds of the original 3,261 determinands were used). Although the EA database[10] contains 24 years of datapoints, the data has been sampled sporadically and with low frequency. Importantly, the relatively few samples showing orthophosphate (2% of 70 million datapoints) do not necessarily represent the patterns in all samples and are likely inadequate for carrying out traditional correlation studies, such as Spearman analysis, to select best features. In addition, this limited dataset did not show clear trends or seasonality, suggesting that fluctuations are likely event driven. The following sections aim to provide a detailed overview of the steps taken for feature selection.

A key outstanding question is how to perform feature selection in a way that identifies representative features for prediction, reduces bias, and captures explained variance – an essential step in AI/ML model development. To address this, we conducted a molecular structure analysis of the chemical determinands to understand the molecular similarities and patterns that can influence Orthophosphate. This analysis utilised Large Language Models (LLM)[11], Deep Learning models and Natural Language Processing (NLP) libraries.

Based on the analysis of molecular similarity using chemical libraries and AI/ML-based models, we have identified several key tools for chemical structural similarities. The models

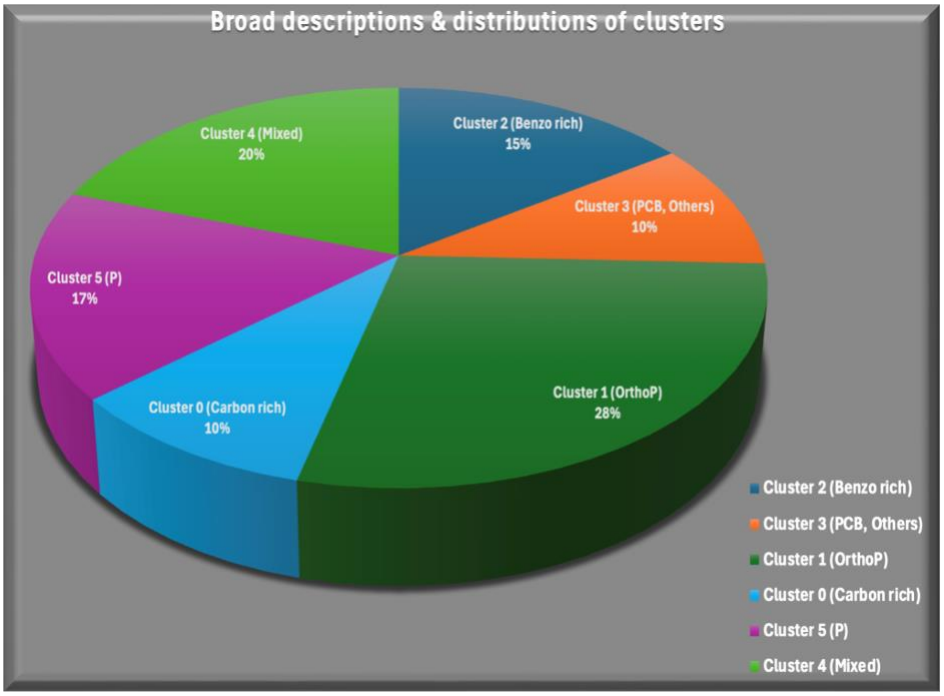
10 Environment Agency's publicly available Water Quality monitoring database

11 LLM is a type of AI/ML model, trained on enormous data, known for their ability to understand and generate contextually relevant text (e.g. Generative Pretrained Transformer - GPT)

used include the deep learning-based transformer model ChemBERTa[12], the open-source ML model RDKit[13] for molecular similarity analysis, NIH[14]-based open-source chemical database PubchemPy[15] and the advanced NLP-based chemical named entity recognition engine (NER[16]) ChemDataExtractor[17] from the University of Cambridge, UK. These tools were deemed suitable for feature analysis of our water quality dataset.

Each of these tools offers distinct strengths: ChemBERTa leverages transformer-based language models to interpret chemical semantics, ChemDataExtractor automates text mining for structured chemical data from literature, PubChemPy provides API access to PubChem's vast chemical database, and RDKit converts molecules into Morgan fingerprints for precise similarity analysis. Combined with K-Means++ for optimal feature clustering, these tools enable efficient molecular similarity studies and feature selection.

We conducted experiments with machine learning tools (mentioned above) and identified six groups of determinands, including two clusters related to phosphorus. Agglomerative analysis on SMILES-converted (Simplified Molecular Input Line Entry System) chemical names revealed molecular similarities, highlighting two dominant orthophosphate related clusters.



12 Transformer model for SMILES-based molecule embedding and actively used in research at Cambridge, Stanford

13 Core library for molecular similarity analysis, extensively used at Oxford, Cambridge, MIT, Stanford and industries (e.g., Novartis, Pfizer)

14 National Institute of Health – a primary agency for the United States of America govt., responsible for biomed and public health research founded in 1887

15 Python wrapper accessing PubChem database which is maintained by National Institute of Health-NIH through the National Center for Biotechnology Information-NCBI, U.S.A

16 Named Entity Recognition is a real-world object that can be uniquely identified by a name like Nitrogen as chemical

17 ChemDataExtractor – a chemistry aware Natural language Processing ML (Machine Learning model) pipeline

Figure 3: Distribution of determinands

Figure 3 above depicts the distribution of determinands (features) into respective groups (clusters). The clusters identified as orthophosphate and phosphorus are particularly significant for our analysis. However, it is worth noting that a determinand, while relatively close but placed in another cluster, may also be considered a distant relative. Multiple experiments were conducted to finalise the list of related features in the latter phase. These two clusters include phosphorus and orthophosphate, respectively, which indicates a logical relationship between them. Each cluster name has been assigned based on broad categorizations.

The following section outlines the processing steps in detail.

4.1. Feature-selection methodology – ChemBERTa

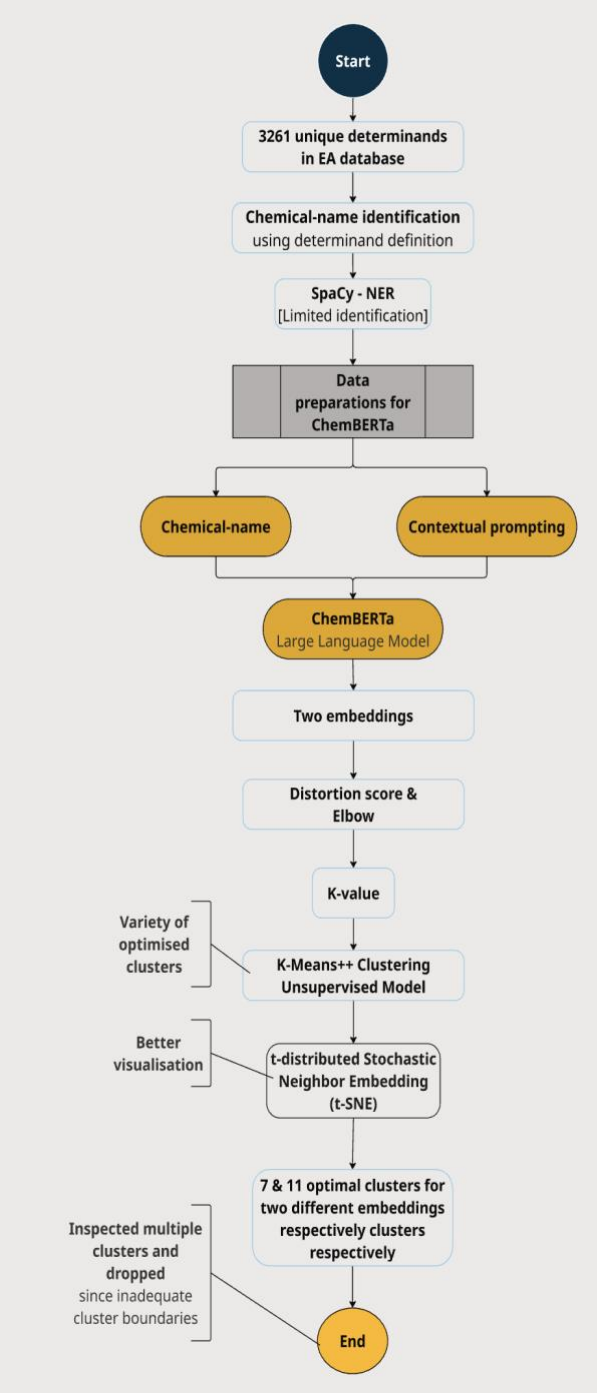
Clustering technique using ChemBERTa with a RoBERTa Tokenizer (Huggingface Weblink)

Summary box:

- **Step1**: Spacy's Named Entity Recognition (NER) was used to identify the presence of chemical names in the determinand definition. **Limitation:** This NER identified around 700 out of 3261 determinands

- **Step2:** Using ChemBERTa with RoBERTa Transformer model, two embeddings were created. One with the NER produced chemical names and another using a contextual embedding

- **Step3:** Using the Distortion score, an Elbow value was calculated to find an optimal k-value for the K-means++ model. Various input values were experimented to find best cluster



ChemBERTa based on RoBERTa model[18] documentation here

18 RoBERTa (short for "Robustly Optimized BERT Approach") is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model, which was developed by researchers at Facebook AI

4.1.1. Analysis and decision Summary:

Overview: The feature-selection pipeline identifies chemical names, processes language, creates embeddings with a transformer model, and uses unsupervised clustering to categorize 2,295 unique chemical determinands.

Methodology: Data from 3,261 unique determinands in the EA database were used. Chemical names were identified using definitions and SpaCy's NER. SpaCy can group determinands by chemical name, but it has a limitation: it grouped 2,627 determinands into one group (Figure 4) while others formed smaller groups (Figure 5).

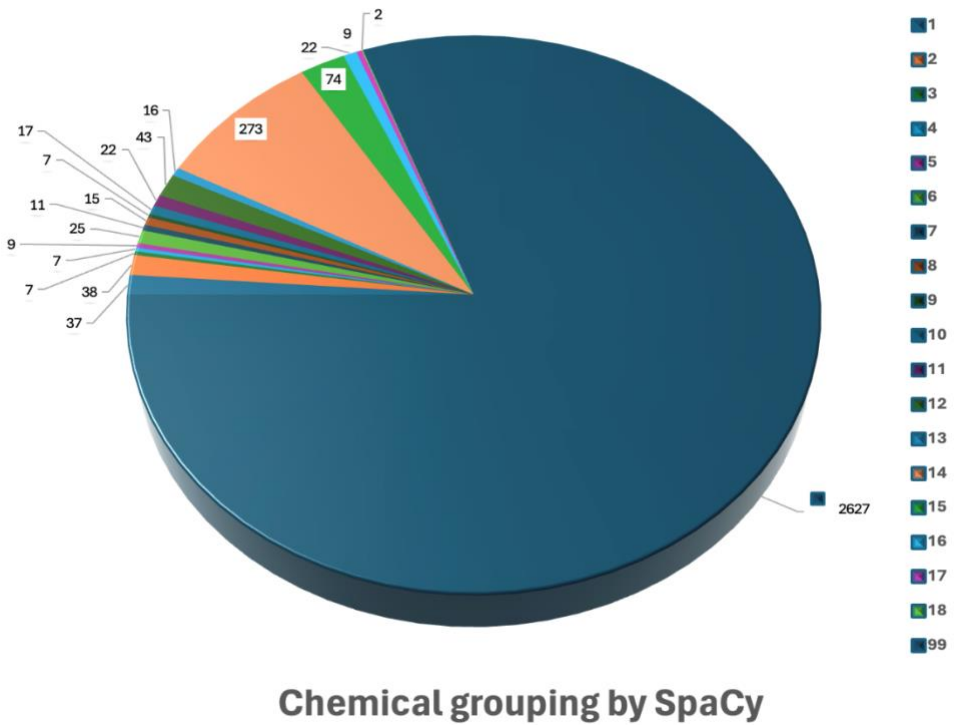


Figure 4: Chemical grouping

The grouping of 80% of determinands into one group makes it inadequate and biased, and therefore we decided to use the advanced method provided by ChemBERTa for further analysis. ChemBERTa generated two complementary embeddings for clustering optimisation using distortion scores and the elbow method to find optimal k-values. K-Means++ clustering produced multiple clusters, which were optimised and visualised with t-SNE. We found 7 and 11 clusters to be optimal.

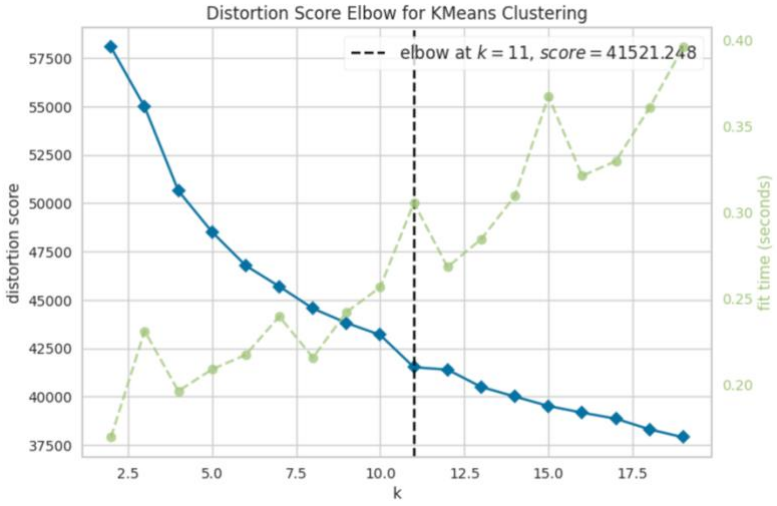


Figure 5: Distortion score Elbow for K-Means clustering

Results: Using the advanced methods provided by ChemBERTa the clustering pipeline processed all 3,261 determinands again, identifying 7 and 11 as optimal clusters (refer to Elbow line in Figure 5). Scatter plot visualisations (see Figure 6 below for 11 clusters) were used for quality checks. Each dot represents a determinand, and the grouped chemicals do not form clear boundaries.

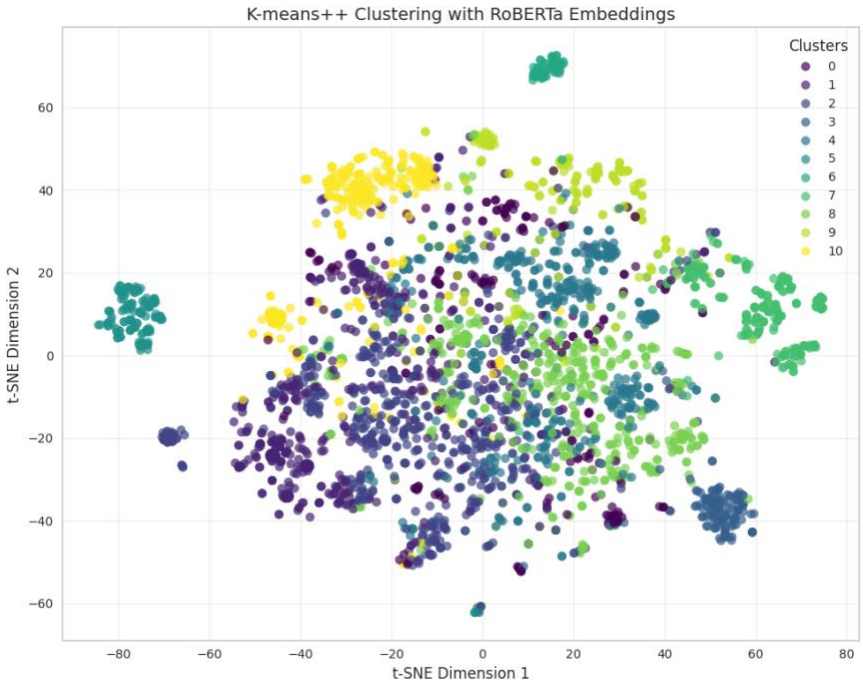


Figure 6: K-means++ clustering with RoBERTa embeddings

Challenges and Limitations: The pipeline organises chemical determinands using NLP and unsupervised learning techniques. ChemBERTa embeddings combined with K-Means++ clustering and t-SNE visualisation provide insights into chemical relationships in environmental analytics. We created two types of embeddings: one using only chemical names for word embeddings resulted in 7 clusters and another using contextual prompting for sentence embeddings resulted in 11 clusters.

However, neither of these approaches gave clear boundaries. These clusters have poorly separated boundaries, which raise a question on how well the model understands the relationships within the given determinands. This uncertainty led to further research for better methods.

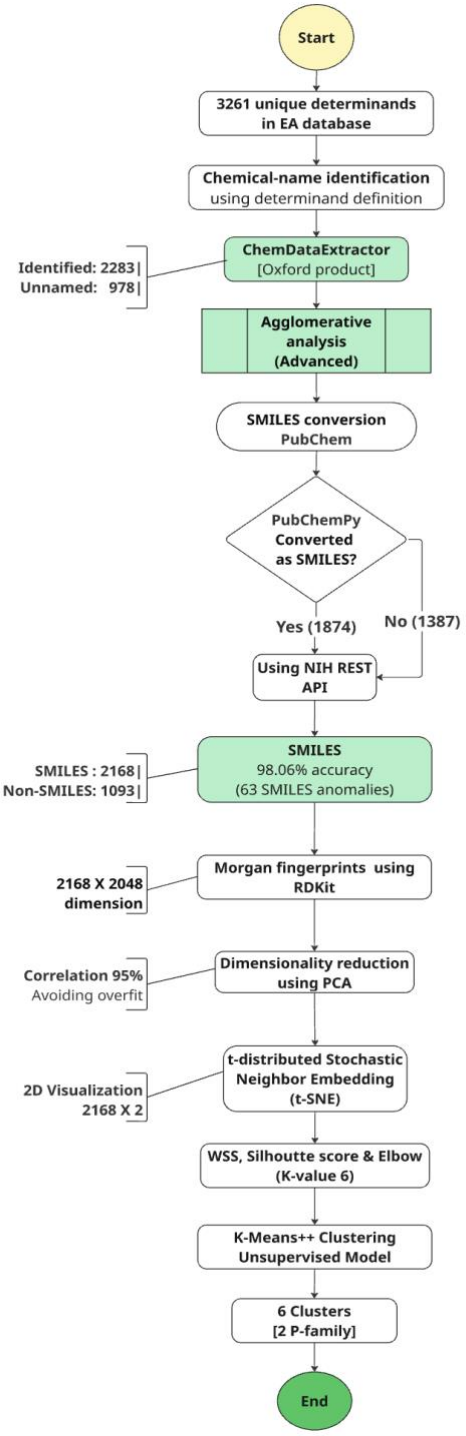
Decision: Future improvement is required. In molecular similarity analysis, it was found that the ChemDataExtractor library performs better in recognising a chemical in a given text, and agglomerative clustering is more advanced in similarity analysis. When combined, both methods yield significant results.

4.2. Feature-selection methodology – Agglomerative clustering

Clustering technique with an advanced agglomerative analysis

Summary box:

- **Step1**: Spacy's Named Entity Recognition (NER) was first used to identify the presence of a chemical name in the determinand definition. Limitation: This NER was able to identify around 700 out of 3,261 determinands
- **Step2**: Using ChemDataExtractor, we overcame this limitation. This library was able to identify 2,283 determinands and left 978 determinands unidentified
- **Step3**: Using PubChemPy with the NIH national chemical library, 2,168 determinands were predicted as SMILES (Simplified Molecular Input Line Entry System) and 1,093 as non-SMILES. Upon further analysis, 63 and 245 (Total: 309) respectively were identified as anomalies, resulting in 91% accuracy in SMILES conversion
- **Step5**: Using RDKit library, these SMILES were converted into Morgan fingerprints, also known as circular or extended-connectivity fingerprints (ECFPs), a type of molecular fingerprint used in cheminformatics to represent the structure of a chemical molecule. Morgan fingerprints are represented in a 2168 X 2048 matrix
- **Step6**: Given the high dimensionality, we chose Principal Component Analysis (PCA) with a 93% correlation for dimensionality reduction. According to AIML industry standards, 90% represents a better fit, and while above 95% is the best fit, it can lead to overfitting. To avoid overfitting, we opted for 93%, providing 356 optimal components
- Using the Stochastic Neighbourhood Embedding algorithm (t-SNE), the PCA outcome was converted into 2 dimensions for better visualization
- **Step7**: The reduced features were fed into K-means++, yielding six optimal groups of determinands based on the Within Cluster Sum of Squares and Silhouette score



River Deep Mountain AI

Footnote references: WCSS[19], Silhouette[20]

Article reference: Molecular Fingerprints are a molecular formula represented in bit form. Click here for related article from ***Novartis Institutes for BioMedical Research*** *lab at London.*

Footnote references: ECFP[21], SMILES[22], Morgan fingerprints[23], ChemDataExtractor & PubChemPy[24]

4.2.1. Analysis and decision summary:

The clustering of the determinands was done using a K-Means++ unsupervised model. The process commenced with an agglomerative analysis, where the determinands were converted into their respective molecular forms in their bit representation. Subsequently, the K-Means++ algorithm was applied to this bit represented form. The detailed procedure is outlined below.

An agglomerative analysis was carried out to improve clustering accuracy. This method is also known as hierarchical clustering algorithm. This is a bottom-up approach as distinct from the other main hierarchical clustering method (known as divisive), which is a top-down approach. This process creates a hierarchical structure of clusters. This concept was applied to the SMILES text (Simplified Molecular Input Line Entry System) to convert it into a Morgan Fingerprint, a standard bit-form representation of molecular structures. SMILES is a text-based notation used in chemical analysis to represent the structure of chemical molecules.

SMILES Example: (Click here for the NIH reference to try a sample conversion)

- Orthophosphate (OrthoP): [O-]P(=O)([O-])[O-]
- Nitrate: [N+](=O)([O-])[O-]
- Nitrite: N(=O)[O-]
- Oxygen: O=O
- Ammonia un-ionised as N: [N]
- Nitrogen: N#N
- BOD (Oxygen Demand): CC1=CC(=C(C=C1OC)C(CN)OC)OC

19 WCSS, also known as inertia, measures clustering efficiency in ML algorithms like K-Means++. It helps determine the optimal number of clusters.

20 Silhouette score is a metric used to evaluate the quality of cluster, how well each datapoint fits within its assigned cluster

21 ECFP Type of Morgan fingerprints that use a specific algorithm to reduce computational cost of generating the fingerprints

22 Simplified Molecular-Input Line Entry System (SMILES) notation encodes chemical structures for computer processing

23 Molecular descriptor represents a molecule as a binary vector, showing the presence or absence of specific substructures within a defined radius around each atom. Sample python implementation from Alcesflight, Oxfordshire – article.

24 Product of Cambridge professor Mathew Swain , state of the art NLP that automatically extracts chemical information from documents | provides PubChemPy an interface to access PubChem allows chemical searches by name, substructure, similarity and retrieval of its properties

More explanations for SMILES by the US Environmental Protection Agency (EPA) can be found here.

Morgan fingerprints are a type of molecular descriptor that represent molecules as a bit string, where each bit indicates the presence or absence of a specific substructure within a defined radius around each atom. These methods are frequently utilised in cheminformatics, similarity searching, and structure-activity relationship studies, which align precisely with the objectives of our feature-selection methodology.

Morgan fingerprints[25] example:

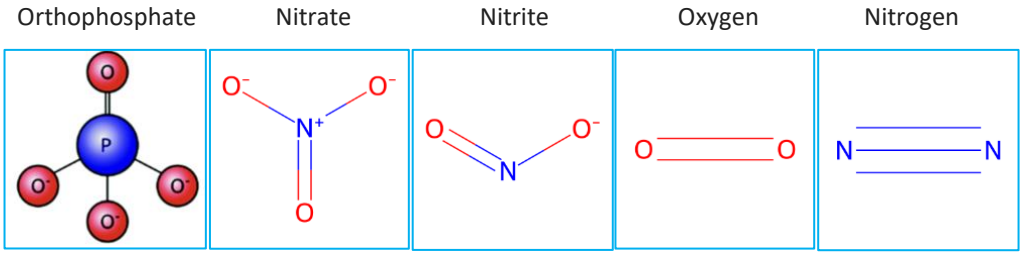


Figure 7: Chemical (molecular) structures used for Morgan fingerprint conversion

Figure 7 represents a chemical in its molecular structural form. Using a 'fingerprint' mechanism, these structures would be encoded in bit form. We utilised the Morgan mechanism, which is primarily used for identifying similar structures. This method aids in our study by allowing us to compare the similarity between molecules. For instance, when calculating the Morgan fingerprints of two molecules with different chemical determinands related to water quality, A and B, if their bit strings exhibit considerable similarity (i.e., many identical bits set to 1), the resulting similarity score between A and B will be high. This indicates a significant degree of structural resemblance between the two molecules. They have demonstrated a correlation and exhibit similar influence when reacting in an aquatic environment. We compare all structurally represented applicable determinands using the Morgan fingerprint bit representation, aided by an unsupervised K-Means++ machine learning model, and group them into optimised clusters. In this study, we have identified 2,168 determinands with a negligible error corresponding to high accuracy, acknowledging the presence of some incorrect identifications, but they are minimal. These determinands are then clustered based on their similarity scores, resulting in six closely grouped clusters.

25 References to journals in similarity search conducted in the past:

A) Click **here** – Cheminformatics journal (2020) about molecular fingerprints. B) Click **here** - Most recent (2024) research conducted in similarity search

The fingerprint conversion process resulted in a 2168 x 2048 matrix for each identified determinand. Subsequently, Principal Component Analysis (PCA) was performed to reduce these matrixes into fewer dimensions. Then, a t-SNE algorithm was applied to further reduce this into a 2-dimensional matrix, which was then supplied to K-Means++ for clustering. This was the final product we used. On top of this, we wanted to select highly influential features out of a correlation coefficient analysis conducted. A key principle in feature selection is to choose the minimal number of representative features. Although we could include all orthophosphate-related clusters (hundreds of features) in model training, we opt to start with a minimal set of features and incorporate the rest only when necessary.

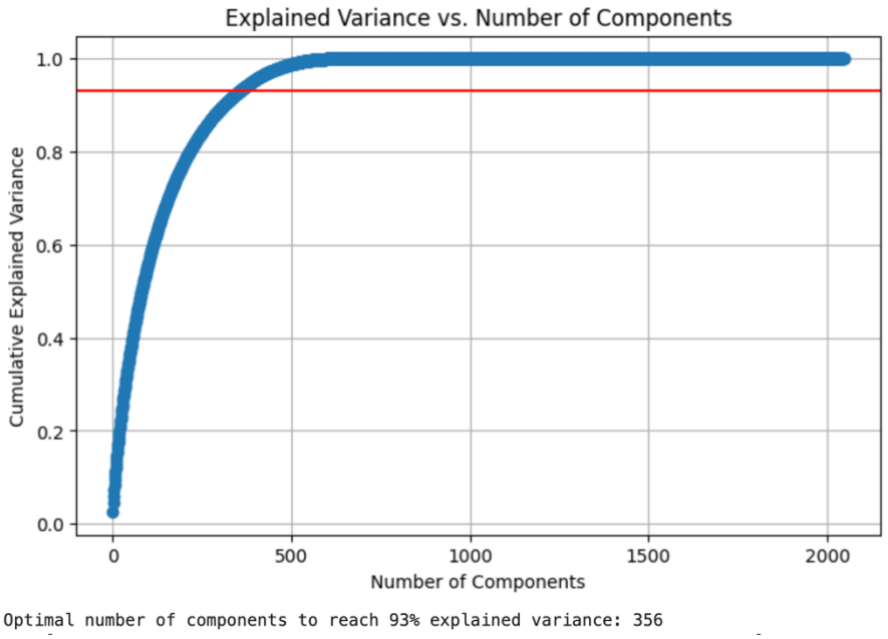


Figure 8: Explained variance vs number of components.

PCA with a 93% explained variance threshold was used for the Morgan Fingerprints. PCA represents the high-dimensional matrix in 356 components (Figure 8), explaining 93% of its variance. It addresses issues from strong correlations.

A correlation of 65% demonstrates a good or moderate relationship, 95% indicates a strong relationship, and 70% to 95% is considered better. We aim to maintain this range to prevent the model from learning features that lead to overfitting.

Inference on the K-Means++ cluster results: We applied PCA for Dimensionality Reduction (356 principal components with 93% variance) on Morgan fingerprints (2168x2948 matrix), followed by K-Means++ clustering. For better visualisation, the PCA output was further reduced to 2D with t-SNE (6 clusters, 2 dominant orthophosphate related clusters), likely representing significant chemical groupings.

Figure 9 illustrates a t-SNE projection of molecular structures clustered into six groups using K-Means++, with two dominant orthophosphate related clusters (clusters 1 and 5), which are suspected to be orthophosphate and phosphate-related compounds. The separation of clusters along t-SNE Components 1 and 2 indicates that molecular structural features can effectively distinguish chemical groups. This could assist in identifying key pollutants or functional patterns in water quality analysis.

The advantage of this approach is its ability to reveal hidden patterns in high-dimensional data. The apparent separation of clusters suggests that the structural similarities and differences are well captured by the model. In this process, dimensionality reduction methods, such as PCA, were implemented before obtaining the visuals using t-SNE to confirm a more reliable interpretation. t-SNE's tendency to emphasise gaps between groups may exaggerate true molecular differences, necessitating validation with actual chemical properties.

Point to note: The PCA and t-SNE analyses in this study are done on a multi-dimensional matrix outcome derived from the Morgan fingerprint conversion. Unlike our usual feature reductions done by PCA methodology, this analysis was not directly applied to the water quality dataset. Instead, it focused solely on the chemical names (determinand definitions containing chemical names), without involving any dataset containing data points. This unique approach only requires the names of the chemicals, not the actual data points. The determinands identified in clusters 1 and 5 were subsequently presented to domain experts for further evaluation.

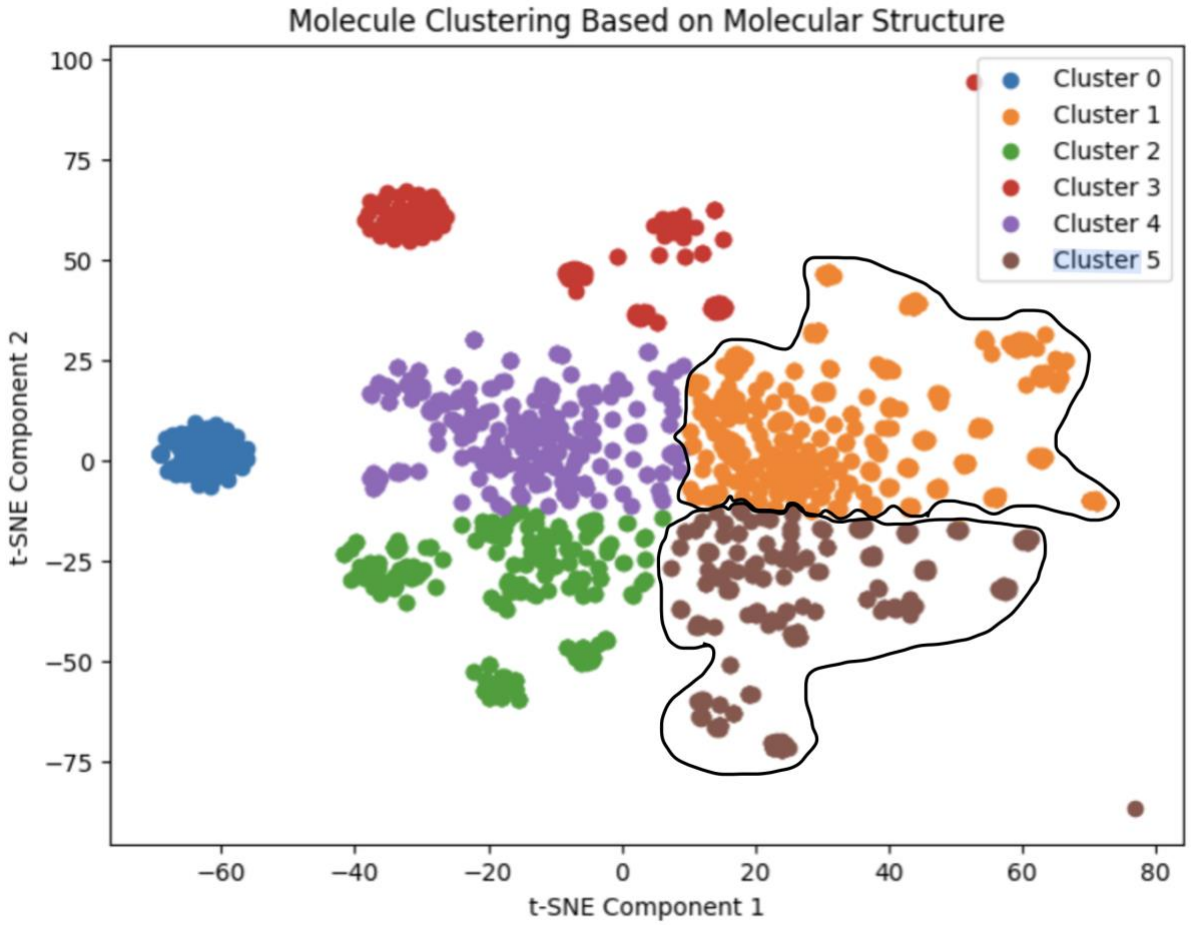


Figure 9: Outcome of the clustering process showing 6 distinct clusters. The dominant orthophosphate-related clusters (clusters 1 and 5) have boundaries drawn around them

As we moved on to the modelling phase, we finalised features from the clustering process and non-chemical data. This was done by collaborating with domain experts and by analysing non-chemical data using Spearman monotonic directional ranking. We decided that this hybrid approach would be beneficial for our model.

We conducted a correlation study to rank the determinands identified within clusters 1 and 5 and in the proximity of these clusters, combined with determinands selected by domain experts. We then validated these findings through multiple experiments, as detailed in the following sections. Ultimately, a total of 45 determinands were selected through evaluating multiple experiments.

4.3. Other feature selection methodologies

Ahead of moving into the modelling phase, we wanted to verify that our more complex agglomerative clustering approach provided better feature selections compared to more traditional and relatively simpler approaches such as Pearson and Spearman correlation analysis.

Given that orthophosphate represents just 2% of 70 million observations, using the Pearson correlation coefficient for feature selection is challenging. This is because the

relatively small orthophosphate dataset is inadequate to explain relationships with other features. Therefore, an optimised non-linear feature selection methodology was implemented. This method comprised of a four-stage intuitive selection process. Despite this adaptation, few relevant features were obtained, which were deemed inadequate for modelling considering the 3,261 features of our water quality dataset. These are mostly attributed to insufficient orthophosphate observations in our dataset.

The initial step was to prepare data prior to analysis. Detailed descriptions of the data preparation process can be found in Section 5 "Feature Engineering", under 'Data exploitation' and 'Final data preparation' topics.

4.3.1. Pearson correlation coefficient method:

We applied the Pearson correlation coefficient technique to a stratified sample collection, but most correlations ranged from -0.000089 to +0.0306 with a few above 0.1, falling very short of the industry average mark of at least 30% for feature selection.

> *Key consideration: This was one of the primary reasons that led us to reject the EA dataset during the Proof of Concept (POC) phase for predicting primary pollutants like phosphate/orthophosphate.*

4.3.2. Optimized non-linear feature selection pipeline:

Backward-feature elimination and forward-feature selection are two prominent techniques for identifying relevant features in high-dimensional datasets.

> ***What is Backward Feature Elimination:*** *It involves considering all available variables initially, calculating the model's performance, eliminating one variable at a time, recalculating the model's performance, and repeating this process until no further variables can be eliminated. The Scikit-learn Recursive Feature Elimination with Cross-Validation (RFECV) tool was utilised to facilitate our evaluation of suitability.*

> ***What is Forward Feature Selection:*** *It begins with a single feature, incrementally adding single features, and repeatedly calculating and comparing performance metrics until all features are included. Due to the non-linear nature of our data, we employed the "SelectKBest" with "f_regression" from the Python ML library for evaluation.*

Our approach integrated the above selection methods to create a four-stage optimised non-linear feature selection pipeline for identifying influential features for the target feature orthophosphate.

> **Stage 1:** This is a noise removal stage; 970 near-constant features are removed using VarianceThreshold (threshold=0.02).

> *Key observation: An extremely low variance threshold (2%) was adopted here based on a rule-of-thumb cut-off. Features where >= 98% of values are identical were removed as they are effectively*

constant for the model's purpose. Slightly higher thresholds resulted in fewer than 9 features being retained.

Stage 2: Fast Correlation Screening employs Spearman ranking (Threshold = 0.29), retaining features with |Spearman correlation ranking| above the threshold. In Spearman analysis, correlations up to 0.3 indicate a weak relationship and are unsuitable for selecting representative features. Correlations above 0.7, however, are ideal for this purpose. This is particularly crucial as orthophosphate constitutes only 2% of the overall population. No features were selected unless this weak ranking (0.3) was maintained. Although this method is not ideal for identifying related features, it provided insight and clear direction for adapting the feature selection strategy.

Stage 3: Optimised using 'mutual_info_regression[26]' to measure the dependency between independent variables to the target variable orthophosphate. The top features that correlate with orthophosphate values, exhibiting at least 30% correlation, are selected.

Stage 4: LightGBM-RFECV. This stage entails Recursive Feature Elimination with Cross-Validation (RFECV) using the LightGBM model as the estimator. This has resulted in zero features selected. However, given that we mandated the retention of the weakly related features (in Stage 2), feature selection using the LightGBM model was disregarded. As a result, we identified nine weakly related features.

Key observation: It is implemented with an adaptive dynamic step size, employing a tailored logic. Instead of eliminating a fixed number of features (e.g., one feature) at each iteration, which would result in numerous iterations for 1000s of features, this approach ensures the removal of 10% of the features per iteration, focusing on those that are preferably less correlated. As the process nears completion, fewer features are eliminated at each step to preserve the most correlated ones.

Outcomes summary: After conducting multiple iterations using these techniques, we identified nine[27] features for orthophosphate. Spearman correlation values up to 0.3 indicate a weak relationship and are not optimal for feature selection, while values of at least 0.7 are suitable for selecting representative features. Without forcing the inclusion of all weakly related features, no features are selected. Consequently, the feature selection using LightGBM (Stage 4) was constrained to retain all features without employing selection, bypassing the necessary feature selection step. We had to choose which weaker features to retain, which is not an advisable method. The table below explains the outcomes of the experiments.

Identifying features with or without LightGBM feature selection	With weak correlation	With optimal correlation
Forced without feature selection	9	0

26 Scikit learn ml library reference here

27 These features are N Oxidised, Nitrite-N, Ammonia(N), Nitrate-N, Temp Water, pH, Alky pH 4.5, BOD ATU, Chloride Ion

Using feature selection	NA	0

Table 1: Traditional correlation method's outcome

LightGBM-RFECV investigates the non-linear relationships and selects representative features accordingly. Feature-selection step cannot be considered complete without this process.

Through these experiments, we were only able to identify a limited number of weak correlations inadequate to build our model. This supported our assumption and gave confidence in the need for our more complex agglomerative clustering approach (detailed above).

4.4. Necessity, uniqueness and significance of clustering:

Clustering ensures that the model prioritises structurally similar determinands (e.g., orthophosphate related determinands), thereby reducing noise from unrelated variables. This approach provides several intangible benefits. Clustering analysis serves as an advanced feature selection methodology, aiding the model in building on structurally similar features. Additionally, we have implemented the traditional correlation approach for the remaining determinands where clustering analysis was not applicable. Consequently, we have adopted a hybrid approach with two main benefits:

Consistency: The model is supplied with a group of water quality (WQ) determinands that are structurally similar to orthophosphate, ensuring a standardised outcome for each prediction. This also addresses potential data drifts, which are critical steps for model maintenance. Correlation, being a mathematical analysis, relies heavily on the observed values (data points) of each determinand and can yield different weightings for different data sets.

Extensibility: Clustering has identified hundreds of orthophosphate-related determinands. As a result, the model can predict any of these determinands with minimal adjustments to the existing code and target determinand, without requiring new exploratory data analysis (EDA) or additional analysis.

The primary distinction between clustering and traditional feature selection methods are:

S.No.	Clustering method	Non-clustering method
1	**Focus:** The analysis focused on the chemical similarity of the determinands, with no data involved.	**Focus:** Analysis was performed on the filtered dataset. Data filtration was carried out with the assistance of domain experts.
2	**Context:** The analysis of molecular similarity was conducted using SMILES and Morgan fingerprint structures.	**Context:** The correlation of the dataset's non-linear nature was analysed.
3	Determinands were categorised into six clusters based on their molecular similarity. We have the flexibility (freedom of choice) to select from	The analysis yielded 9 determinands from the thousands of total determinands.

	hundreds of P-related clusters as well as any closely related clusters. Ultimately, a total of 45 determinands were selected through evaluating multiple experiments.	
4	**Tools:** LLMs, Segmentation MLs were used in this analysis. SpaCy (NER based), ChemBERTa-RoBERTa, ChemDataExtractor, RDKit, PCA, t-SNE, NIH-PubChemPy were used.	**Tools:** Optimised non-linear feature selection pipeline using LightGBM model, Scikit-learn's mutual-info - f_regression, Spearman's ranking, Backward feature elimination, Forward feature selection. Pearson's correlation coefficients analysed
5	**Technique:** Closely positioned orthophosphate clusters were analysed as well.	**Technique:** Correlation considered above 30%, using an adaptive selection method.
6	**Based-on:** Analysis conducted on two-thirds of the determinands molecular similarity. Outcome was experimented on the data points.	**Based-on:** An analysis was conducted on over 30+ million data points, with a 2% representation for orthophosphate.

Table 2: Distinction between clustering and traditional feature selection methods.

Hybrid approach undertaken: A comprehensive hybrid approach was employed, which included incorporating feedback from domain experts, utilising advanced feature selection methods such as agglomerative clustering, traditional correlation-based analysis, combined with multiple experiments, together resulted in finalising representative features.

We combined the features from all methodologies and finalised them for our Open Orthophosphate Model. See the 'Final features' under 'Final Model' section (Section 8) for complete list of features.

- 45 features were identified by combining the features from the cluster analysis [28] and domain expert[29] consultation. Please refer to Table 9 for a comprehensive list of determinands. Some examples include:
    - Features identified by cluster analysis are *ammoniacal nitrogen, nitrogen total oxide, nitrate, nitrite, alkalinity to pH 4.5 as CaCO3, BOD 5Day ATU, chloride, calcium, magnesium, hardness, silica, potassium, fluoride, copper, and zinc* etc., including the Section-82 determinands ammonia (NH_3) un-ionized, dissolved oxygen, oxygen dissolved %saturation and oxygen dissolved (Laboratory) as O_2.

28 Feature selection using chemical cluster analysis

29 Features considered in consultation with water domain (Refer **Final-feature** section for more details)

- Features identified by consulting with domain experts are conductivity at 25 °C and weather – precipitation and air temperature including the Section-82 determinands water temperature and pH.

- Seven temporal and spatial features were added to these 45.

5. Feature Engineering

After choosing features, we moved on to the next phase for preparing the data (refer to node 2 in Diagram 1). This involves analysing and engineering the features to identify significant patterns in the dataset for the final model. First, it is important to have well-organised data, which is achieved in the data exploitation step.

Data exploitation: The water quality data in the EA archive follow a unidimensional format, i.e., one reading per observation per determinand. Understanding the nature and distribution of the data is crucial for identifying potential features and data transformations accordingly. More so, any machine learning model understands the data better if the data are represented in a two-dimensional format. Hence, we transposed the data into a two-dimensional, model consumable format.

Sample Date	determinand notation	determinand name	determinand definition	result	unit_name	SamplingPoint notation	SamplingPoint name
05/06/2024	0111	Ammonia(N)	Ammoniacal Nitrogen as N	10.4	mg/l	AN-011396	Green House
05/06/2024	0116	N Oxidised	Nitrogen, Total Oxidised as N	1	mg/l	AN-011396	Green House
05/06/2024	0117	Nitrate-N	Nitrate as N	2.59	mg/l	AN-011396	Green House
05/06/2024	0118	Nitrite-N	Nitrite as N	0.028	mg/l	AN-011396	Green House
05/06/2024	0119	NH3 un-ion	Ammonia un-ionised as N	0.0013	mg/l	AN-011396	Green House
05/06/2024	0180	Orthophospht	Orthophosphate, reactive as P	9.71	mg/l	AN-011396	Green House
05/06/2024	0192	Phosphate	Phosphate :- {TIP}	0.128	mg/l	AN-011396	Green House
06/06/2024	0111	Ammonia(N)	Ammoniacal Nitrogen as N	7.4	mg/l	AN-011496	Home Nursing
06/06/2024	0116	N Oxidised	Nitrogen, Total Oxidised as N	1.1	mg/l	AN-011496	Home Nursing
06/06/2024	0117	Nitrate-N	Nitrate as N	3.59	mg/l	AN-011496	Home Nursing
06/06/2024	0118	Nitrite-N	Nitrite as N	0.082	mg/l	AN-011496	Home Nursing
06/06/2024	0119	NH3 un-ion	Ammonia un-ionised as N	0.0031	mg/l	AN-011496	Home Nursing
06/06/2024	0180	Orthophospht	Orthophosphate, reactive as P	9.17	mg/l	AN-011496	Home Nursing
06/06/2024	0192	Phosphate	Phosphate :- {TIP}	0.281	mg/l	AN-011496	Home Nursing

Table 3: Illustrating a Uni-dimensional with sample values from EA's water quality dataset

Sample Date	determinand notation	determinand name	unit_name	0111	0116	0117	0118	0119	0180	0192
05/06/2024	AN-011396	Green House	mg/l	10.4	1	2.59	0.028	0.0013	9.71	0.128
06/06/2024	AN-011496	Home Nursing	mg/l	7.4	1.1	3.59	0.082	0.0031	9.17	0.281

Table 4: Illustrating Two-dimensional transposed format of the above Uni-dimensional sample values

Final data preparation with domain experts: Filters were applied to the dataset based on domain expert suggestions. The dataset comprises 3,261 unique determinands collected from over 64,000 sampling points, with a total sample size of 68.47 million. Domain experts recommended focusing solely on surface water samples for modelling, reducing the sample size to 33.57 million, with 1,897 determinands.

With the support of domain experts, certain determinands were identified as not appropriate for use, such as "type of flow description" (determinand notation: 3267) with 193,445 records

and “beach signage confirmation” (determinand notation: 3724) with 224 records, and Chemical Oxygen Demand[30] (determinand notations: 0090, 0091, 0092, 8080) with 152,821 records and so these were removed.

After applying the filters to the 68.47 million (as per recommendations as stated above), the final sample size is 33.42 million, with 1,891 determinands from 23,000 sampling points in England’s WQ monitoring dataset. This dataset can potentially be used to model and predict orthophosphate concentrations at the 64,000 sampling points, taking into account the limitations in the data available and in the model. The following sections explain the methodologies used. The 1,891 determinands from the cleansed ‘surface water' samples are a subset of the total 3,261 determinands. Following the dataset cleansing, we need to identify the orthophosphate influencers within this cleansed dataset comprised of 1,891 determinands for our model-building exercise.

Key consideration: During modelling, determinands measured fewer than 100 times over the 24 years were excluded. Consequently, 905 out of 1,891 determinands were removed.

5.1. Trend and Seasonality:

Trend refers to the long-term, general direction of the data, indicating an upward or downward movement. Seasonality refers to recurring patterns or cycles in the data that occur at regular intervals, often tied to calendar events or natural cycles. Understanding both helps in forecasting and analysing data over time.

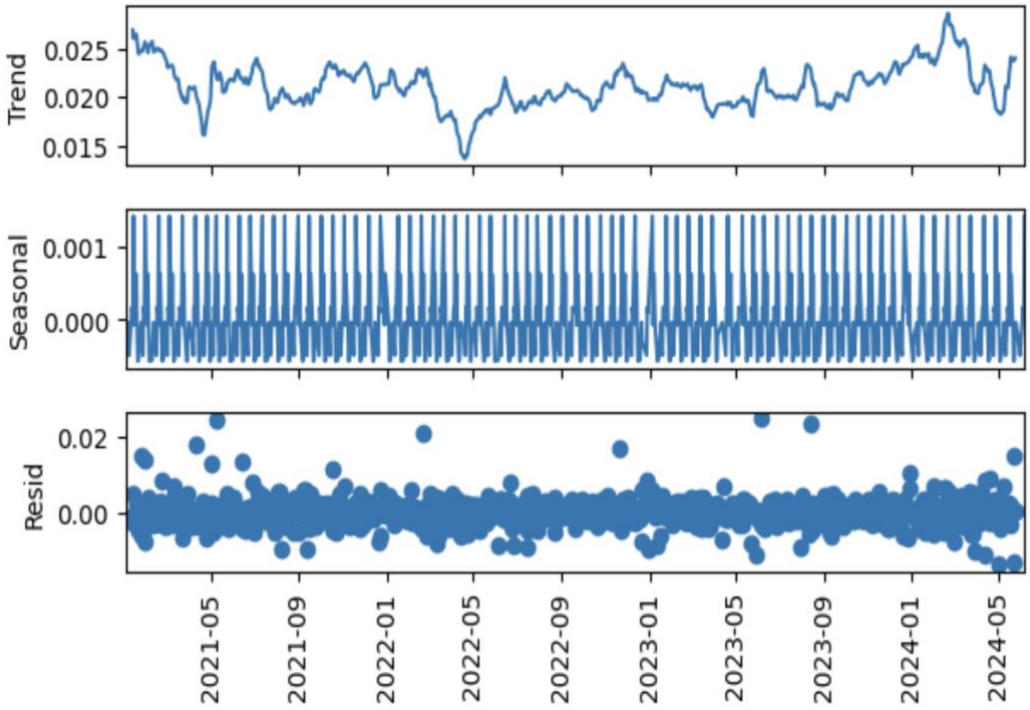


Figure 10: Trend and Seasonality analysis

30 Determinands such as “Chemical Oxygen Demand, Filtered” (COD), “Chemical Oxygen Demand, Settled” (COD), “Chemical Oxygen Demand” (COD), “Chemical Oxygen Demand, Leachable: Dry Wt” (COD)

Three years of orthophosphate data from 2021 to 2024 were taken for this analysis. Data during 2020 did not bring any insights, since the measurements were negligible or not taken due to covid lockdown, hence considered post covid years data.

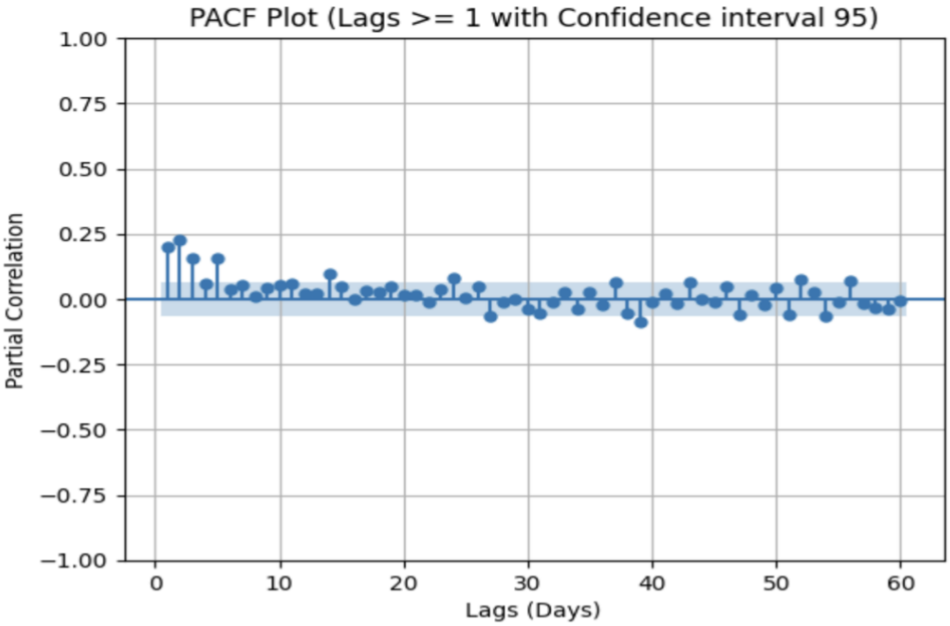
Trend: *The target variable orthophosphate has no strong or long-term trends over a year (May 2021 to May 2024) data.*

Seasonality: *Clear annual cycle with consistent peaks and troughs. Repeating spikes/dip present at fixed intervals (e.g., May-July & January-March respectively). As a next step, further analysis to confirm inference, ACF (Autocorrelation Function) / PACF (Partial Autocorrelation Function) based analysis is employed.*

Residual: *Noted several +0.02 mg/L spikes, indicating short-term contamination. Adding factors like rainfall, industrial discharge, overflow alerts, and agricultural activities to the model may enhance its accuracy.*

5.2. ACF & PACF:

Autocorrelation and Partial autocorrelation indicate whether today's orthophosphate measured value depends on previous day's observations or any day in the past (for example, a day last week). If they show significant correlation, then the data are not independent; otherwise, they are. As shown in the graph below, the data suggests little or no time dependency, particularly after 2-3 days. This might be due to the low frequency of samples in the dataset, which is often only every other month. For the same reason, timeseries modelling is not applicable for this type of dataset.



Checking for autocorrelation using ACF and PACF for January 2021 to May 2024

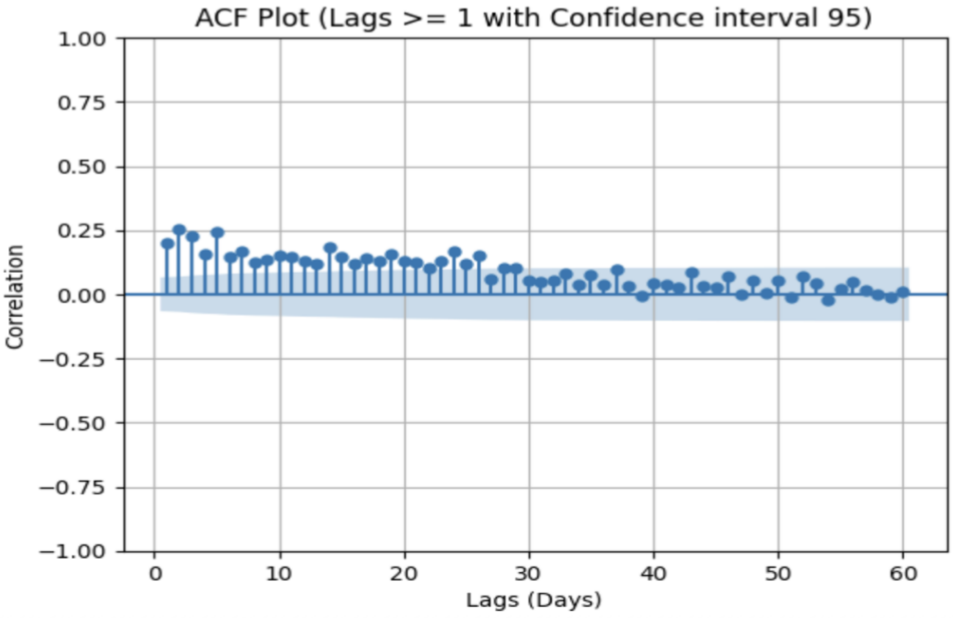


Figure 11: PACF and ACF plots to analyse trend and seasonality

Key observations:

Metric	*Observation*	*Implication*
ACF	***Trend:*** *The autocorrelation function (ACF) values exhibit a slow decay pattern indicates a persistent trend in the data.* ***Seasonality:*** *There are no distinct seasonal spikes or repetitions at consistent intervals. No clear seasonality.* *Non-Stationarity: Failure to quickly drop to zero confirms a non-stationary behaviour.*	*Non-stationary*

PACF	*Key pattern: Sharp cut-off after initial lags There is a significant correlation at lag 1, with no notable spikes at other lags. This indicates that there are no direct correlations beyond lag 1. Consequently, any features created, such as rolling mean or X-past days exploitation, are unlikely to reveal a pattern.* *Here we present the aggregated average orthophosphate values collected over three years from various sampling points across England. Typically, the orthophosphate measurements are taken at intervals of 1-2 months per sampling location. However, for the aggregated three-year data, the most common intervals are 1, 7, and 28 days.*	*-NA-*

Table 5: ACF and PACF Observations

5.3. Stationarity & non-stationarity:

During the above ACF, PACF analysis (Trend and/or Seasonality assessments), we identified non-stationarity in orthophosphate concentrations. Data are said to be "non-stationary" if it shows a trend and has an autocorrelation structure over time. Here is the hypothesis test to confirm:

1. ***A Null hypothesis (H0):*** *The series is non-stationary if it has some time-dependent structure and does not have constant variance over time*
2. ***Alternate hypothesis (HA):*** *The series is stationary*

With these hypotheses, we applied the Ad Fuller (ADF) Test for stationarity. We chose three samples: one was a specific sampling point (Series 1), and the other two were randomly selected multiple sampling points (Series 2 & Series 3).

Series (data)	ADF Statistic	p-Value	Critical values at 1%, 5%, 10% significance
Series 1: One sampling point in Northwest region of England 'NW-88001164'	-3.418 (Number of observations: 1,185)	0.0103	- 01%: -3.44 - 05%: -2.86 - 10%: -2.57
Series 2: Stratified sample of 18,693 sampling point locations across England	-3.45 (Number of observations: 136,626)	0.012	- 01%: -3.6 - 05%: -2.89 - 10%: -2.58
Series 3: Stratified sample of 23,238 sampling point locations across England	-46.05 (Number of observations: 573,830)	0.0	- 01%: -3.43 - 05%: -2.86 - 10%: -2.56

Table 6: Ad Fuller (ADF) analysis

The ADF[31] statistics for both Series 1 and 2 were more negative than the 5% critical value, but less than the 1% value. The p-values for both were also less than the 5%

31 ADF – Ad Fuller test is used to determine if a data series is stationary. It tests a null hypothesis and alternate hypothesis to be true or not using ADF test statistic and p-value. Based on the interpretation a right model is chosen to get reliable results.

value but greater than the 1% value. This confirms a weak stationarity in the data. Series 3 was not analysed due to the large negative ADF statistic.

To address this weak stationarity, we developed three temporal features that capture seasonal and trend components (see Section 5.4). Given these results, the model should not need continuous re-training, unless accuracy diminishes (e.g., with warmer climate data in 2025).

Please note, our final model was subsequently trained using data from the Environment Agency (EA) spanning January 2000 to May 2024. This allowed for the inclusion of an additional year's worth of data, to address any potential decrease in accuracy during validation phase in June 2025. Future iterations may benefit from incorporating engineered features that capture more temporal behaviours (e.g. cyclical encoding features for seasonality).

5.4. Feature creation:

This process involves deriving new features from existing ones, such as combining multiple features or creating interaction terms to reveal specific patterns at the sampling point level. For temporal features, we included season[32], day of week, and month of year. Additionally, we incorporated spatial features, such as the sampling point name and its BNG[33] coordinates, to access data specific to sampling point aggregations and catchment areas.

5.5. Label encoding:

This process involves converting categorical features into numerical representations. Dummy variables are created for metric units of the determinands. Label encoding is applied to ComplianceSample, purpose_name, samplingPoint_notation, longitude, and latitude. These steps help provide more spatial and temporal patterns of the sampling points/catchments.

5.6. Scaling:

We assessed the impact of feature scaling, but it made no difference to model accuracy. Feature scaling is not required for LightGBM as it assigns weights rather than relying on raw data, therefore we did not scale data.

6. Model selection:

To build robust models and ensure accuracy, we conducted extensive exploratory data analysis (EDA), to try and find spatiotemporal patterns such as seasonal trends or catchment-specific behaviour (Diagram 1, node 3). Advanced techniques like agglomerative clustering and K-Means++ segmentation identified groups of determinands influencing phosphate levels. Our optimised feature selection pipeline, using Spearman ranking, LGMBRegressor

32 Four Seasons: Spring (March, April, May), Summer (June, July, August), Autumn (September, October, November), and Winter (December, January, February)

33 The British National Grid (BNG) is a coordinate system used for mapping and geographic referencing in Great Britain

with Recursive Feature Elimination Cross Validation (RFECV), and mutual_info regression, streamlined feature selection. Additionally, experiments using selected data based on domain expert feedback, with various sets of features, helped finalise the representative features best for prediction model.

The modelling phase tested a suite of algorithms:

- Statistical linear regression model for baseline insights
- Tree-based approaches focused on producing consistent outcomes each time the model were applied (Random Forest, XGBoost) to capture non-linear relationships
- Tree-based with gradient boosting approaches focused on improving accuracy each time the model was applied (LightGBM) to capture non-linear relationships
- Artificial neural networks (ANN) for high-dimensional pattern recognition

The following steps were undertaken when selecting the final model

- Selected six models, the first five of which are spatiotemporal aware models that learn patterns of spatial and temporal dependencies
    i. Random Forest
    ii. XGBoost
    iii. LightGBM
    iv. LightGBM Hypertuned
    v. ANN
    vi. Traditional linear regression
- Input from domain experts.
- Conducted multiple experiments and evaluated model performance.
- Created train, test and validation datasets for evaluating model performances.
- Produced four different data-cuts to analyse model performances:
    i. National dataset comprised of 171 determinands with 8500 rows each. One targeting phosphate and another targeting orthophosphate prediction.
    ii. Spatiotemporal dataset comprised of 60 features with 46.7 million rows, including samples from five sample categories: RIVER / RUNNING SURFACE WATER, ESTUARINE WATER, SEA WATER, CANAL WATER, and POND / LAKE / RESERVOIR WATER.
    iii. Spatiotemporal dataset comprised of 45 determinands with 33.7 million rows, including Material Types[34] for 'RIVER / RUNNING SURFACE WATER' with Reactive monitoring Purpose Types
    iv. Final dataset: Spatiotemporal dataset comprised of over 45 determinands with 31.7 million rows, including Material Types for 'RIVER / RUNNING SURFACE WATER' without Reactive monitoring.

34 In EA water quality dataset, the "material type" refers to the specific type of sampled material being analysed. This can include various water bodies like rivers, lakes, or coastal areas, as well as different types of water samples (e.g., surface water, groundwater, or wastewater).

- Conducted more than 70 experiments using combinations of the above models, data-cuts and metrices.

Broad categories of metrics were employed to evaluate the performances of the six selected models (outlined above). Their performances, denoted as M1, M2, and M3, (Table 7a) are evaluated as follows:

M1 metrics	RMSE, R Squared, Q Metrics
M2 metrics	RMSE, Normalised RMSE, R Squared, Adjusted R Squared, Q Metrics
M3 metrics	RMSE, Normalised RMSE, MSE, MAE, Responsible AI

Table 7a: Three sets of metrics for performance evaluation

7. Experiments

The objective of conducting the experiment is to decide on a final model architecture that is consistent and reliable. The experiments consisted of

I) Phosphate and orthophosphate predictions using National Dataset. The 'National Dataset' mentioned in this context refers to the aggregation of a 70 million dataset into one observation per day per determinand at a national level.
II) Orthophosphate predictions using a data cut specific to few sample types
III) Orthophosphate predictions using a data cut specific to Surface samples
IV) Orthophosphate predictions using a data cut specific to Surface samples without reactive monitoring data

with the metrics M1, M2, and M3 (see Table 7a) used to determine the model's performance. Train, Test, and Valid are the three sets of data used for model training, testing, and validation, respectively.

Types of Experiments						
Broad categories	**Prediction models**	**Data cuts**	**Dataset volume**	**Material Types[1]**	**Purpose Type[2]**	**Metrics Tracked**
I	All six models	National dataset with a predictor 'Phosphate'	8500 rows	All available types as per EA's WQ Dataset	All types applicable to the Material types	M1, M2
	All six models	National dataset with a predictor 'Orthophosphate'	8500 rows	All available types as per EA's WQ Dataset	All types applicable to the Material types	M1, M2

II	All six models	Spatio-Temporal dataset	46.7 million	RIVER / RUNNING SURFACE WATER, ESTUARINE WATER, SEA WATER, CANAL WATER, POND / LAKE / RESERVOIR WATER	All types applicable to the Material types	M2
III	LightGBM	Spatio-Temporal dataset	33.7 million	RIVER / RUNNING SURFACE WATER	All types applicable to the Material types	M3
IV	LightGBM	Spatio-Temporal dataset	31.7 million	RIVER / RUNNING SURFACE WATER	All but without Reactive monitoring	M3
1. EA's List of Sampled Material Types (URL: https://environment.data.gov.uk/water-quality/def/sampled-material-types.html?_sort=label)						
2. EA's List of Purpose Types (URL: https://environment.data.gov.uk/water-quality/def/purposes.html?_sort=label)						

Table 7b: Broad categories of the Experiments

Prediction models are distinct models evaluated using these different types of datasets, totalling five models with one (LightGBM) of the existing models *hypertuned*[35]. Optimisation models are the list of models used to select optimised features from the 3,261 WQ features in the EA dataset. Approximately 70 experiments were executed to identify the best-performing model, which was then considered the final model for further use. Table 8a, Table 8b and Table 8c summarise these experiments, their variations and results. Our analysis utilised a total of six models: five standard models and one hypertuned LightGBM model, combined to form a comprehensive stack. We categorised the data into four main groups: the complete set of EA samples for a national level aggregation and three unique spatiotemporal data sets as mentioned above in Table 7b. To evaluate performance, we used various metrics as detailed in below tables (refer to Tables 8a, 8b and 8c).

We conducted targeted experiments on phosphates and orthophosphate using four distinct data cuts and evaluating performance with three sets of model metrics. This approach resulted in approximately 6 x 4 x 3 experiments, plus a few more using selective sample types.

	Train	Test	Valid	Dataset Type	Dataset Volume	Metrics Tracked
All Params (170, 150, 18, 58)	A	B	-			

35 Hyperparameters are configuration settings chosen before training the model that control a machine learning model learning. Click here for more details. The process of using these parameters to achieve optimal performance is known as hypertuning.

	Only Sec-82 Params	-	C	-			
Six Models =>	Phosphate	A & B	A & C	-	National Dataset	~8500	M1
	Orthophosphate	A & B	A & C	-	National Dataset	~8500	M2
M1 - Metrics	RMSE, R2, Q Metrics [Training data Variance, Bias, Prediction Variance]						
M2 - Metrics	RMSE, NRMSE, R2, Adjusted R2, Q Metrics [Training data Variance, Bias, Prediction Variance]						
Prediction Models	Statistical Regression, Random Forest, XGBoost, Artificial Neural Network, Light Gradiant Boosting Model, Light GBM Hypertuned						
Optimization Models	SpaCy, chemBERTa, RoBERTa, PubChem, K-Means++ | Feature reduction tools like PCA & t-SNE						

Table 8a: Experiments Summary – Phosphate and Orthophosphate models using national dataset

		Train	Test	Valid	Dataset Type	Dataset Volume	Metrics Tracked
	All Params (60)	A	B	-			
	Only Sec-82 Params	-	C	-			
Light GBM Model ->	Orthophosphate (With 5 Material Types)	A & B	A & C	-	Spatio Temporal	46.7 M	M2
M2 - Metrics	RMSE, NRMSE, R2, Adjusted R2, Q Metrics [Training data Variance, Bias, Prediction Variance]						
Prediction Models	Random Forest, XGBoost, Light GBM, LightGBM Hypertuned						
Optimization Models	RDKit, Morgan Fingerprints SMILES, PubChem, ChemDataExtractor, K-Means++ | Feature reduction tools like PCA & t-SNE | Pearson & Spearman Correlation Coefficient						

Table 8b: Experiments Summary –Orthophosphate models for five sample types (refer model selection section)

		Train	Test	Valid	Dataset Type	Dataset Volume	Metrics Tracked
	All Params (45)	A	-	-			
	Only Sec-82 Params	-	B	C			
LightGBM	Orthophosphate (With R.Mon & 1 Material Type	A & B	A & B	A & B & C	Spatio Temporal	33.7 M	M3
	Orthophosphate (Without R.Mon & 1 Material Type	A & B	A & B	A & B & C	Spatio Temporal	31.7 M	M3
M3 - Metrics	NRMSE, MSE, MAE & **Responsible AI (SHAP)**						
Prediction Models	Light GBM						
Optimization Models	RDKit, Morgan Fingerprints SMILES, PubChem, ChemDataExtractor, K-Means++ | Feature reduction tools like PCA & t-SNE | Optimised feature selection pipeline using mutual_info_regression and LightGBM Regressor						

Table 8c: Experiments Summary – Orthophosphate models for surface samples

7.1. Experiments & Prediction models' performance evaluations

Our work initially focused on phosphate, rather than orthophosphate, hence the first set of experiments (models 8a-1, 8a-2 and 8a-3) focus on phosphate. After internal discussion with domain experts, it was decided to pursue orthophosphate instead, which is represented in models 8a-4, 8a-5, 8b-1, 8b-2, 8c-1 and 8c-2.

8a-1. Phosphate model using national dataset with 171 features

Performance Metrics	Regular Metrics				Q-Metrics		
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed)
Linear Reg	0.35		-0.15		-15.40%	58%	6.00%
Random Forest	0.4		-0.5		-49.90%	97%	0.10%
XGBoost	0.4		-0.51		-50.70%	91%	3.50%
ANN	0.36		-0.21		-20.70%	73%	5.70%
LightGBM	0.39		-0.46		-46.10%	94%	0.60%
LightGBM_HyperTuned	0.33		0.01		1.30%	11%	0.10%

Inference: LightGBM with Hypertuned model performed better than other models with 11% bias and 0.33 mg/l RMSE. It has no negative R^2 and explained variance during training. However, this model explains very little variance of the input determinands (dependent variables). Other models reported negative explained variances.

8a-2. Phosphate model using national dataset with 156 features

Performance Metrics	Regular Metrics				Q-Metrics		
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed)
Linear Reg	0.35		-0.15		-15.50%	61.00%	1.20%
Random Forest	0.4		-0.5		-50.40%	98.00%	0.10%
XGBoost	0.4		-0.48		-47.70%	92.00%	2.80%
ANN	0.38		-0.25		-25.30%	70.00%	0.50%
LightGBM	0.41		-0.46		-46.10%	95.00%	0.30%
LightGBM_HyperTuned	0.33		0.01		1.30%	11.30%	0.10%

Inference: LightGBM with Hypertuned model performed better than other models with 11.3% bias. It has no negative R^2 and explained variance during training. It has very little explained variance. Other models reported negative explained variances with high bias.

8a-3. Phosphate model using national dataset with 18 features

Performance Metrics	Regular Metrics				Q-Metrics		
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed)
Linear Reg	0.27		0.17		17.10%	-2.00%	46.20%
Random Forest	0.24		0.36		35.80%	-0.60%	61.80%
XGBoost	0.23		0.4		39.80%	-0.30%	61.30%
ANN	-		-		-	-	-
LightGBM	0.23		0.4		42.00%	0.20%	57.30%
LightGBM_HyperTuned	0.23		0.44		43.70%	0.00%	39.30%

Inference: We removed negative explained variance (EV) with the help of hybrid feature selection explained Section 4.4 and the models showed better goodness of fit. Light GBM gives a stable result comparatively. The ANN model has recorded lower bias.

8a-4. Orthophosphate model using national dataset with 58 features including 20 Section 82 and 7 easily accessible (Orthophosphate outliers removed using domain expert recommended 0.5 mg/l as cap)

Performance Metrics	Regular Metrics				Q-Metrics		
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed)
Linear Reg	1707.95	127554.14%	-137071420	-137085292.8	-13707142090%	27.686	13707128866.00%
Random Forest	0.08	5.97%	0.67	0.67	67%	-0.02	65.00%
XGBoost	0.11	8.22%	0.42	0.42	42%	0.003	36.00%
ANN	-	-	-	-	-	-	-
LightGBM	0.12	8.96%	0.38	0.38	39%	0.003	32.00%
LightGBM_HyperTuned	0.1	7.47%	0.5	0.5	49%	0.004	39.00%

Inference: Notice that the orthophosphate model reported better EVs for training and predictions, and positive R^2 and Adjusted R^2 values. This seemed a better goodness of fit. Model is improving.

8a-5. Orthophosphate model using national dataset with 18 features

Performance Metrics	Regular Metrics				Q-Metrics		
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed)
Linear Reg	0.1	13.77%	-8.47	-10.84	-847%	3.607	0.00%
Random Forest	0.08	11.02%	-4.19	-5.49	-419%	-2.537	0.00%
XGBoost	0.1	13.77%	-7.97	-10.21	-797%	-3.499	0.00%
ANN	-	-	-	-	-	-	-
LightGBM	0.19	26.16%	-32.6	-41	-3260%	-7.077	0.00%
LightGBM_HyperTuned	0.04	5.51%	-0.31	-0.64	-31%	-0.803	35.70%

Inference: Notice that R^2, adjusted R^2 and explained variances were reported negative. This means that model is poorly fitting, and prediction is obtained just using the mean value

8b-1. Orthophosphate model using spatiotemporal dataset (46.7 M) with 60 features as well as 15 additional engineered features without reactive monitoring data

All Independent Variables are used for Training & Prediction	Regular Metrics				Q-Metrics			Residual Errors			
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed	MSE	MAE	MAPE	
Linear Reg	0.05	10.00%	0.42	0.42	0.41	-0.008	0.42	0	0.02		TRAIN & TEST
XGBoost	0.04	8.00%	0.64	0.64	0.64	-0.008	0.57	0	0.01		
LightGBM	0.03	6.00%	0.68	0.68	0.68	-0.007	0.65	0	0.01		
LightGBM_HyperTuned	0.03	6.00%	0.68	0.68	0.68	0.03	0.6	0	0.01		

8b-2. Orthophosphate model using spatiotemporal dataset (46.7 M) with 60 features as well as 15 additional engineered features with reactive monitoring data

All 60 Independent Variables are used for Training. Only 18 Spatio+Temporal+Sec82 used for Prediction	Regular Metrics				Q-Metrics			Residual Errors			
	RMSE	NRMSE	R2	Adjusted R2	Explained Variance Train	Bias	Variance Ratio (Pred/Observed	MSE	MAE	MAPE	
Linear Reg	0.05	10.00%	0.28	0.28	0.28	0.3	0.14	0	0.02		TRAIN & TEST
XGBoost	0.04	8.00%	0.45	0.45	0.45	0.3	0.28	0	0.01		
LightGBM	0.04	8.00%	0.46	0.46	0.46	0.31	0.29	0	0.01		
LightGBM_HyperTuned	0.04	8.00%	0.44	0.44	0.44	0.36	0.27	0	0.01		

Inferences for both above: These two experiments utilised four distinct models, with Light GBM offering consistent results. Domain expert feedback led to a test with and without reactive monitoring data, showing no significant difference in metrics (RMSE, Normalised RMSE and MAE). Reactive monitoring data was included in the final model

Each experiment utilised a distinct number of features based on the hybrid feature selection methods discussed previously.

8. Final Model

In this context the final model is the one for production after evaluating all feature selection, engineering, and tests conducted using exploited data through the execution of various models. With our dataset, models predict non-zero when the real number is NULL, making R^2 and adjusted R^2 report negative scores, which is not useful.

While the Q metrics explain the variability of the supplied 60 independent determinands using the datapoints of the training dataset relative to the target determinand of the model (orthophosphate), it is important to note that we have intentionally exploited the data to detect minute differences in catchment patterns. This understanding adds limited context to the model's variability. Therefore, we decided not to discard them from our evaluation.

The MSE, MAE, and RMSE metrics continue to provide meaningful insights during both Train-Test and Ground Truth validation phases. Therefore, we have opted to assess the model's performance solely using these metrics.

Final features: After conducting feature selection and feature engineering analysis using the methodologies described above, combined with multiple experiments, we found the 45 features shown in Table 9 as the most representative features for predicting orthophosphate levels. These experiments were applied to obtain predictions for all 64,000 sampling points across England catchments. Notably, the model uses orthophosphate data available from approximately 20,000 sampling points in the training dataset and is capable of making predictions for the test dataset with sampling points unmonitored for orthophosphate.

	Definitions as per EA's WQ dataset			OFS -> Optimized feature selection pipeline (non-cluster approach)				
Sl. No.	Determinand Notation	Determinand Name	Determinand Description	FS by Clustering	FS by OFS-pipeline	Confirmed by SME	Confirmed by experiments	Section-82
1	0116	N Oxidised	Nitrogen, Total Oxidised as N	X	X	-	X	-
2	0111	Ammonia(N)	Ammoniacal Nitrogen as N	X	X	-	X	-
3	0117	Nitrate-N	Nitrate as N	X	X	-	X	-
4	0118	Nitrite-N	Nitrite as N	X	X	-	X	-
5	0119	NH3 un-ion	Ammonia un-ionised as N	X	-	-	X	Yes
6	9924	Oxygen Diss	Oxygen, Dissolved as O2	X	X	-	X	Yes
7	0162	Alky pH 4.5	Alkalinity to pH 4.5 as CaCO3	X	X	-	X	-
8	0085	BOD ATU	BOD : 5 Day ATU	X	X	-	X	-
9	0135	Sld Sus@105C	Solids, Suspended at 105 C	-	X	-	X	-
10	0172	Chloride Ion	Chloride	X	X	-	X	-
11	0241	Calcium - Ca	Calcium	X	-	-	X	-
12	0237	Magnesium-Mg	Magnesium	X	-	-	X	-
13	0158	Hardness	Hardness, Total as CaCO3	X	-	-	X	-
14	0182	SiO2 Rv	Silica, reactive as SiO2	X	-	-	X	-
15	0348	Phosphorus-P	Phosphorus, Total as P	X	-	-	X	-
16	0211	Potassium- K	Potassium	X	-	-	X	-
17	3683	N Inorganic	Nitrogen, Total Inorganic : (Calculated)	X	-	-	X	-
18	0192	Phosphate	Phosphate :- {TIP}	X	-	-	X	-
19	9686	Nitrogen - N	Nitrogen, Total as N	X	-	-	X	-
20	0114	N-Kjeldahl	Nitrogen, Kjeldahl as N	X	-	-	X	-
21	0113	N Organic	Nitrogen, Organic as N	X	-	-	X	-
22	0076	Temp Water	Temperature of Water	-	X	X	X	Yes
23	0301	C - Org Filt	Carbon, Organic, Dissolved as C :- {DOC}	X	-	-	X	-
24	0061	pH	pH	-	X	X	X	Yes
25	9901	O Diss %sat	Oxygen, Dissolved, % Saturation	X	-	X	X	Yes
26	0183	Sulphate SO4	Sulphate as SO4	-	-	-	X	-
27	0143	Sld NV@500C	Solids, non-volatile at 500 C	-	-	-	X	-
28	0461	DtrgtAncSyn	Detergents, Anionic	-	-	-	X	-
29	0207	Sodium - Na	Sodium	X	-	-	X	-
30	0463	Dtrgt NncSyn	Detergents, Non-ionic	-	-	-	X	-
31	0177	Fluoride - F	Fluoride	X	-	-	X	-
32	0077	Cond @ 25C	Conductivity at 25 C	-	-	X	X	-
33	6450	Cu Filtered	Copper, Dissolved	X	-	-	X	-
34	6455	Zinc - as Zn	Zinc	X	-	-	X	-
35	0749	Phenols Mono	Phenols : Monohydric as Phenol	-	-	-	X	-
36	0209	K- Filtered	Potassium, Dissolved	X	-	-	X	-
37	0205	Na- Filtered	Sodium, Dissolved	X	-	-	X	-
38	9856	OrthophsFilt	Orthophosphate, Dissolved	X	-	-	X	-
39	7859	SO4dis	Sulphate, Dissolved as SO4	X	-	-	X	-
40	1183	WethPresPrec	Weather : Precipitation	-	-	X	X	-
41	0239	Ca Filtered	Calcium, Dissolved	X	-	-	X	-
42	0175	Cyanide - CN	Cyanide as CN	X	-	-	X	-
43	0235	Mg Filtered	Magnesium, Dissolved	X	-	-	X	-
44	1181	WethPresTemp	Weather : Temperature	-	-	X	X	-
45	0082	O Dissolved	Oxygen, Dissolved : (Laboratory) as O2	X	-	-	X	Yes

Table 9: List of finalised features for orthophosphate model

The final model incorporated these 45 features and spatial characteristics such as label-encoded sampling point notation, latitude and longitude (derived from British National Grid coordinates), and temporal patterns such as day of the week, month of the year, and seasons.

Running LightGBM prediction model utilising these finalised features, we have obtained promising results, providing confidence in our approach:

8c-1. Orthophosphate model using spatiotemporal dataset with 45 features including 6 Section 82 determinands.

Dataset Volume (With Reactive Monitoring) : Data Volume 33.7 Million						TRAIN & TEST
OrthoP prediction	Regular Metrics		Residual Errors			
	RMSE	NRMSE	MSE	MAE	MAPE	
LightGBM	0.04	8.00%	0.00123	0.01		

8c-2. Orthophosphate model validating with 11 features (listed in Table 10) including 5 Section 82 determinands

Dataset Volume (With Reactive Monitoring observations) : 190+ Predictions						VALIDATE (Using ground truth data)
OrthoP prediction with Ground Truth Validation	Regular Metrics		Residual Errors			
	RMSE	NRMSE	MSE	MAE	MAPE	
LightGBM	0.03	7.89%	0.00088	0.02		

8c-1 shows the training and testing performance results of the model using the 45 features, resulting in a very low MSE of 0.00123. This trained model then was utilised to generate predictions for a validation data set on the specific monitoring point (Dearness upstream of Priest Burn, Sampling Point NE-44400163). The performance results of this are shown in 8c-2, where residual errors (e.g. 0.02 MAE and 0.00088 MSE) indicate a strong alignment (goodness of fit[36]) with the observed data.

36 Good fit refers that the model performs well both training and unseen data

Determinand notation	Determinand definition
0076	Temperature of Water
0077	Conductivity at 25 C
0111	Ammoniacal Nitrogen as N
0116	Nitrogen, Total Oxidised as N
0117	Nitrate as N
0118	Nitrite as N
0119	Ammonia un-ionised as N
0162	Alkalinity to pH 4.5 as CaCO3
0180	Orthophosphate, reactive as P
9901	Oxygen, Dissolved, % Saturation
9924	Oxygen, Dissolved as O2

Table 10: Set of 11 features of the validation catchment

9. Final metrics

The metrics MSE, MAE, RMSE and NRMSE and SHAP (Responsible AI framework) and Beeswarm plots are used to evaluate the performance of the final first iteration of the model. Explanations of SHAP and Beeswarm plots are below:

9.1. SHAP metric (Responsible AI framework)

SHAP - Shapley Additive exPlanations – is a method that explains prediction values by the contribution of each feature. These plots are part of the Responsible AI framework, serving to evaluate the model's performance. Metrics like MSE, MAE, or RMSE can often be seen as black box, and SHAP provides clarity in this context.

Figure 12 shows a SHAP force plot, which illustrates how the model derives a specific prediction, showcasing the impact of individual features. Features are displayed at the bottom of the plot, prefixed with the standard determinand notation from the EA's water quality dataset. The base value represents the average or mean value, while the bold value signifies the predicted value.

In the plot, red indicates features that push the prediction towards a higher orthophosphate value, whereas blue represents features that pull the prediction towards a lower value. The size and direction of each impact are significant. We have selected predictions with high and low variations compared to their base values. This visual representation helps identify which features influenced the prediction to be higher or lower.

It is important to note that this plot does not reflect the actual conditions of the catchment or the features influencing the prediction. Instead, it explains how specific values contribute to a prediction. For example, the N_oxidised value of 5.16 might be an outlier or an unusual value. A water domain expert can determine if this is abnormal, and if so, the higher variation in the prediction might be disregarded.

While it is impractical to investigate every prediction, we have provided the source code in the GitHub to analyse suspected incorrect or optimal predictions and understand the model's behaviour. Alternatively, the model's overall behaviour can be visualized using the beeswarm plot in Figure 13.

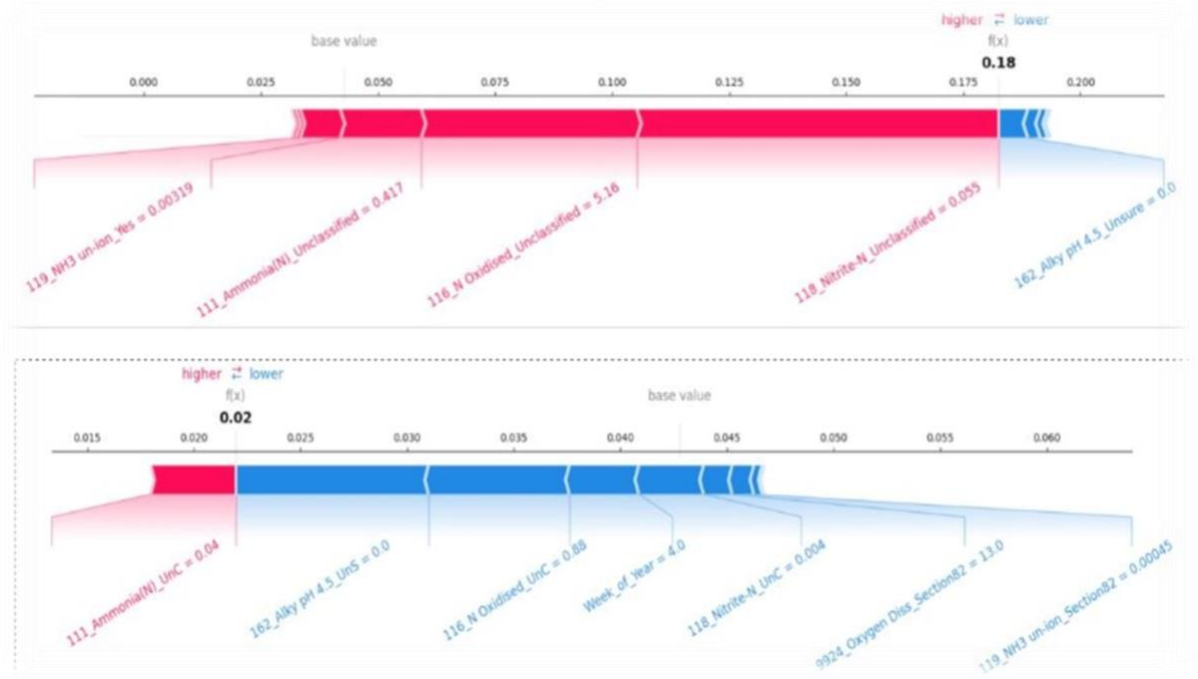


Figure 12: SHAP plots for large (top plot) and small (bottom plot) variations predicted by the model

SHAP plots mean that stakeholders can see how a prediction was made as regulators like the EA demand transparency. Every prediction can therefore be audited – from data sources to weighting factors. This is not a black box; it is a decision-support tool built for compliance.

9.2. Beeswarm plot

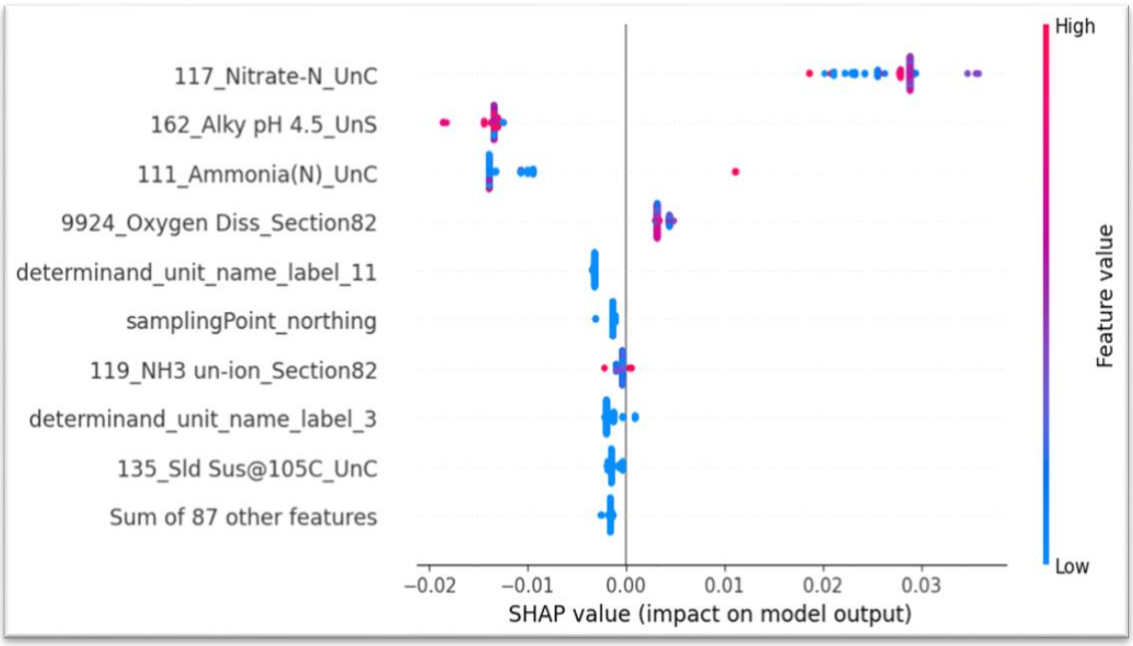


Figure 13: SHAP Beeswarm for the entire 190+ predictions

The beeswarm plot (Figure 13) visualises SHAP values for open orthophosphate predictions, showing each feature's impact on the model. Features such as "Nitrate-N_UnC," "Alky pH 4.5_UnS," and "Oxygen Diss_Section82" are on the y-axis, while SHAP values range from -0.02 to 0.03 on the x-axis. Each dot represents a data point's feature contribution. Aggregated[37] contributors represent minor impacts. The colour gradient shows feature values, with red as high and blue as low. Positive SHAP values increase predictions, negative values decrease them. For example, blue dots with positive SHAP values mean that the low values of that feature increase the value of the predictions. This plot identifies key drivers, such as "Oxygen Diss_Section82" with significant positive influence, and others like "NH3 un-ion_Section82" with mixed effects.

10. Model performance visualization:

The model forecasts orthophosphate levels at the sampling point level to align with EA monitoring standards. For visualisation purposes, the model metrics have been aggregated to catchment level. There are 139 hydrological units (mostly catchments) in the EA's hydrological boundary definition. Each of these catchments contain multiple sampling points. Our predictions based on sampling points for orthophosphate were then aggregated using the EA's defined boundaries for each catchment. The flow diagram below illustrates how the model's performance was assessed and reported spatially.

37 Represents the sum of 87 additional features, specifically created features, offering context for the metric units.

EA monitors water quality at sampling points. For orthophosphate, approximately 18,000 to 20,000 sampling points are monitored across England. Overall, there are 64,000 sampling points that include over 3,261 determinants being observed. By using the orthophosphate measurements from less than one third of the total sampling points, predictions can be made for all sampling points, including those not monitored for orthophosphate, through the open orthophosphate model.

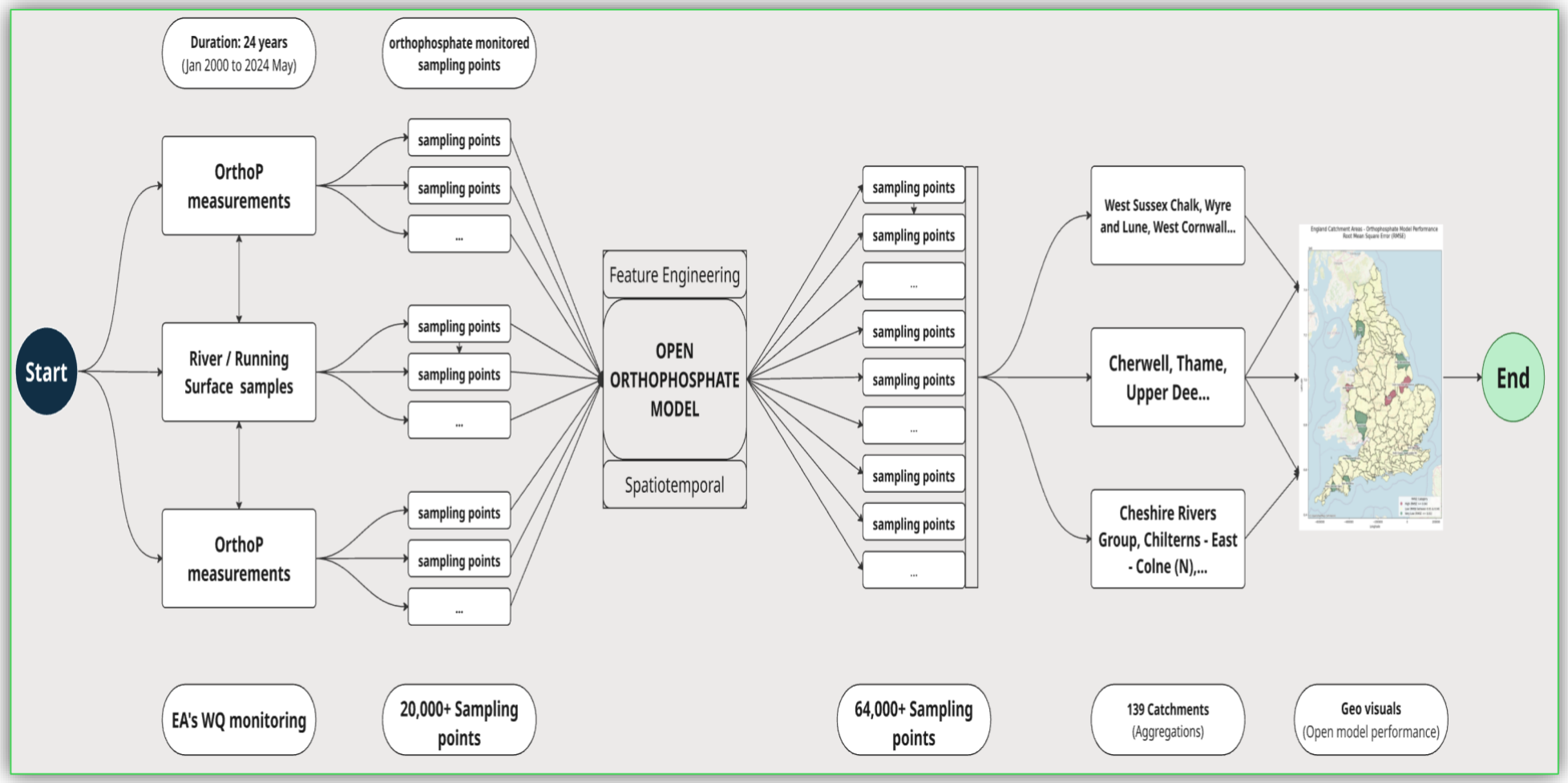


Diagram 2: Aggregation of predicted RMSE summary into Geo visuals for England

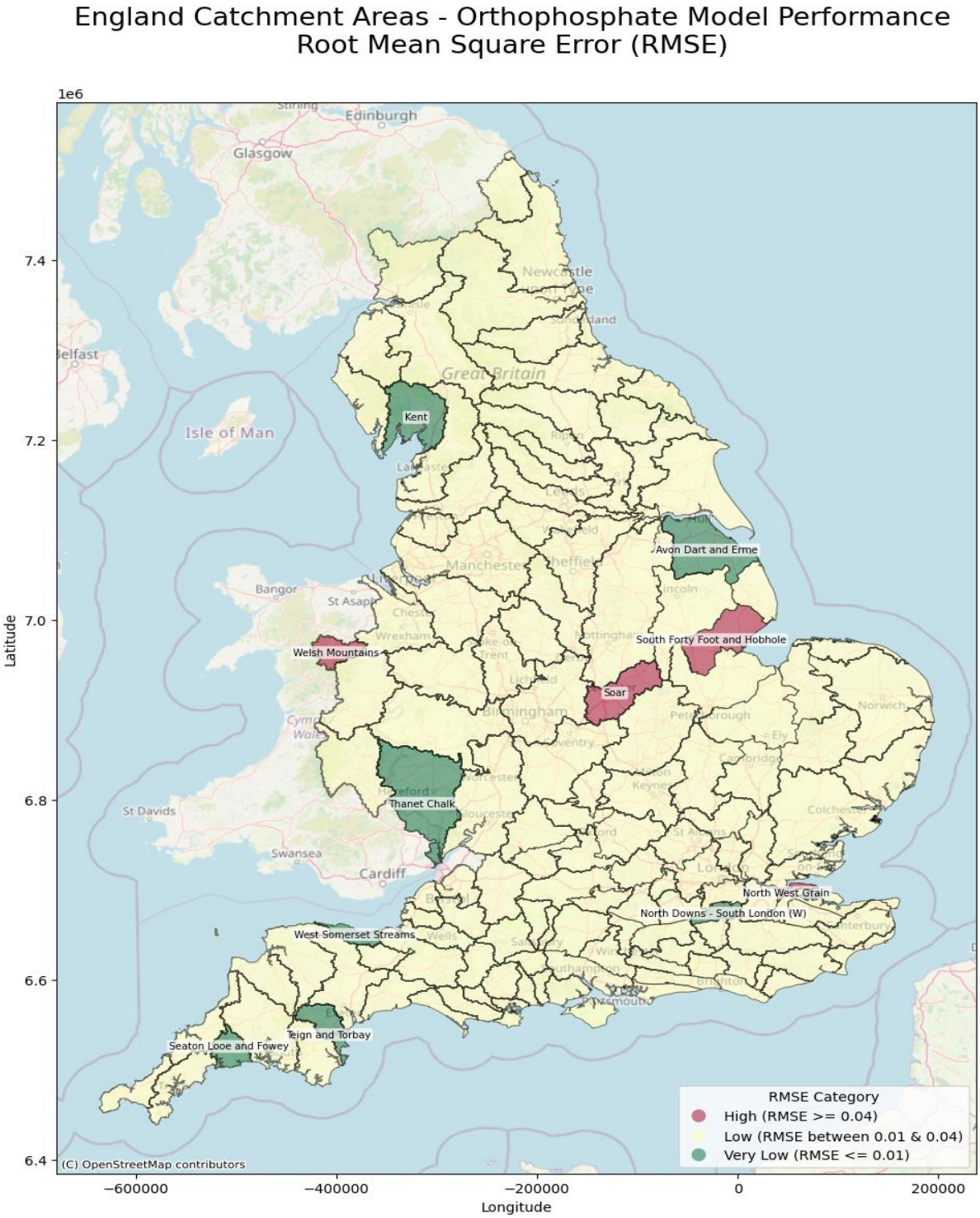


Figure 14: Visualising model performance (RMSE) across catchments (EA Hydrological Boundaries) in England

The Model's RMSE for the year 2023 ranges from 0.01 to 0.06 mg/L. From the multiple experiments we have conducted, a Light GBM model with RMSE of 0.04 mg/L has been concluded as good fit.

The normalised RMSE (NRMSE) is obtained by dividing RMSE by the highest orthophosphate value in our data, which is 0.5 mg/l. The 0.04 mg/l thus gives us an NRMSE of 8%. Given that the AIML standard considers anything below 15% acceptable, our NRMSE of 8% is quite low.

RMSE category	RMSE Range	Catchments	NRMSE Range
Very low	<=0.01	7	<=2%
Low	0.01 - 0.04	127	2% - 8%
High	0.04 - 0.06	4	8% - 12%

Table 12: RMSE and NRMSE Categories

Inference: The model's highest error remains well within the AIML industry best standards, (NRMSE of 8%) with most catchments showing lower errors (refer to Table 12 and Figure 14) indicating the model is predicting with acceptable accuracy.
However, it must be noted the threshold for WFD Good Status for reactive phosphorus can be as low as 0.03 mg/l for some rivers, which is comparable to the RMSE for some monitoring points.

Scatter plot visualisation:

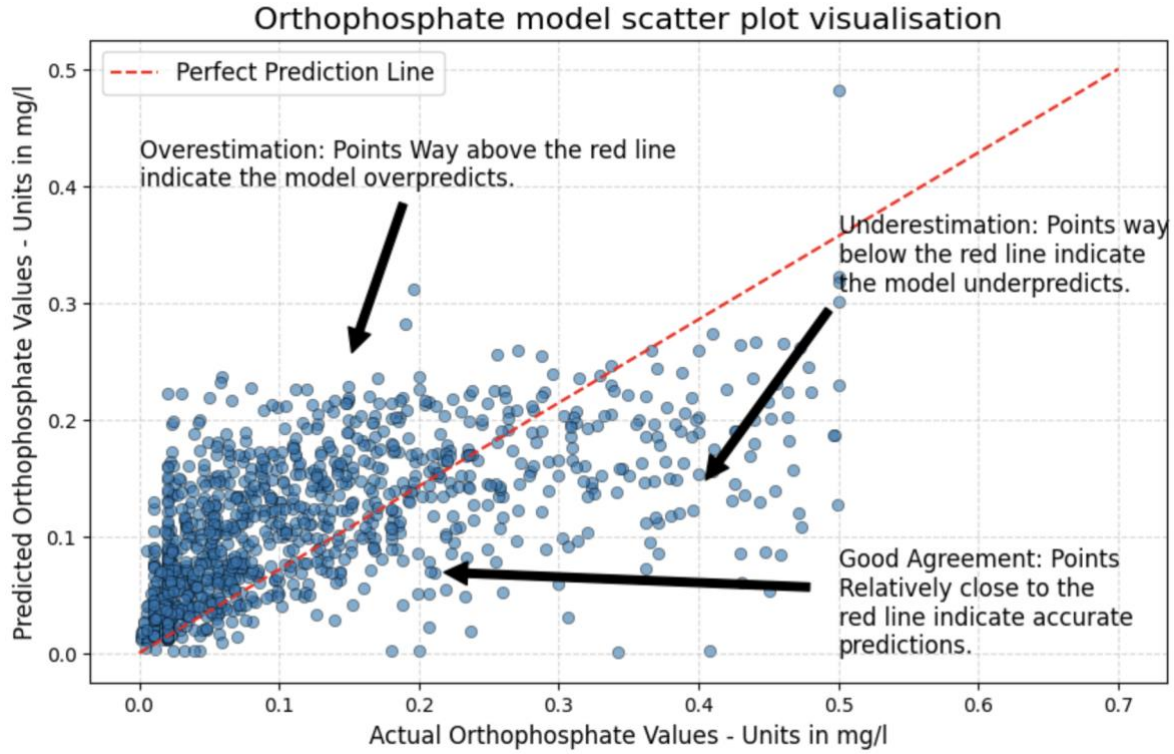


Figure 15a: Orthophosphate measures: Actual vs Predicted (multiple sites)

Figure 15a above depicts the predictive outcomes for unknown values, utilised to evaluate the performance of the trained model. This evaluation employed a 10% test split to ensure the robustness and accuracy of the model's predictions. For enhanced visualisation, a subset of these predicted values has been used.

Yielding overall insights from open orthophosphate model:

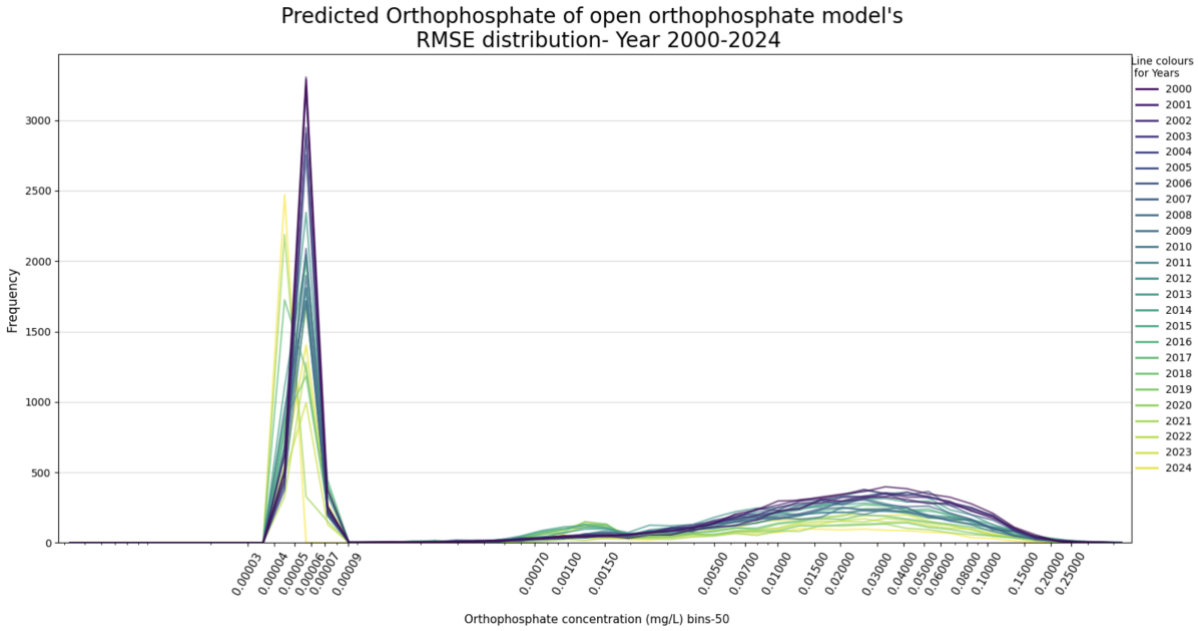


Figure 15b: RMSE (residual error) distribution for the predicted values vs observed values using Train-Test data set

Figure 15b illustrates the performance of the trained open orthophosphate model. The model exhibits high accuracy for typical orthophosphate concentrations, as evidenced by most prediction errors clustering around a near-zero RMSE of 0.00005 mg/l. This clustering of lower measures, which form a majority in the training data set (refer to Figure 2), highlights the robustness of the model. However, the presence of a pronounced right-skewed tail in the error distribution, with RMSE values reaching up to 0.04 mg/l, indicates an opportunity for improvement.

11. Conclusion

The modelling approach has proved the concept of using the optimal set of representative features (45 in total) in monitored water quality data to infer orthophosphate concentration. Future model validation will focus on across the full range of catchments in England, including predominantly urban catchments characterised by runoff and high sewage inputs and predominantly agricultural catchments where diffuse sources dominate. To build public trust, we are open-sourcing core algorithms on GitHub.

The model, and its associated performance metrics, represent the first iteration of our ML-model addressing the challenges mentioned above. As River Deep Mountain AI proceeds into phase 3, these models will be developed further and refined. The results presented here are therefore preliminary and should be considered as such.

The open phosphate model is designed to predict orthophosphate levels for specific sampling points. In England, there are 139 catchments as defined by the Environment Agency's hydrology boundary. Each catchment may have thousands of sampling points. The training code was designed to uncover patterns specific to individual sampling points, thereby minimising the risk of overlooking nuances at the catchment level in predictions. The model comprises distinct training and validation code sets. The validation code, which is publicly available, utilises the trained model stored in a pickle file. The model was trained on a dataset split into 90% for training and 10% for validation, based on a stratified sampling method rather than a fixed date-based approach. This technique allows the model to examine a certain percentage of the volume for each sampling point's data observations, making it fully scalable to all sampling points and catchments across England. The model has multiple potential applications, some of which are described below.

Potential applications of the Orthophosphate Prediction Model:

- The model can predict orthophosphate levels from January 2000 to May 2024, retrospectively, using the existing EA water quality datasets and no additional inputs.
- The model can predict orthophosphate levels at a new sampling point location, even if this location has never been monitored in the past 24 years. There are two methods to achieve this:
    - **Inclusion of Minimal Measured Values:** If actual measured values for the new location are not available, synthesised values can be used. These values are then included in the training data to retrain the model. Note that the training code may require slight modifications to accommodate these time lags in the training and test data.
    - **Utilisation of Nearby Existing Sampling Point:** Consider whether any nearby sampling point from the training data is representative of the desired location and if so, use its input values to make predictions with the trained model.

  By employing these approaches, we can estimate orthophosphate levels even in new monitoring locations, although the confidence in these predictions would be lower.
- The model can be applied to predict orthophosphate levels from alternative datasets, such as newly collected water quality data. This will provide stakeholders with a supplementary tool capable of inferring orthophosphate fluctuations using readily available WQ data, such as real-time section 82 datasets. In phase 3 of the project, we will continue to explore and evaluate this usage of our model. To explore this please see section 16 on validation.
- To generate retrospective predictions for specific past dates without modifying the existing code, utilize the provided model source code. Ensure you supply the data for the relevant past time period and adhere to the input schema detailed in the GitHub documentation to achieve the published prediction accuracy.
- The training and test code can be adjusted to predict future values, such as orthophosphate levels for the next six months at a specific site in England, by

changing the lag value. For example, in forecasting models, we adjust time-lags to simulate future predicted values. We can apply a similar technique here, even for a specific site if necessary. The code modification will need to accommodate the adjustment of appropriate time-lags according to the date intervals for the orthophosphate present in the dataset. This requires minor code modifications.

- According to EA's Water Quality data archive, there are over 64,000 sampling points across England where water quality is being monitored. The model was trained using orthophosphate data collected from over 18,000 of these sampling points but can be used for both those points where orthophosphate is measured and also the 44,000 sampling points where it is not. The model's predictions have shown high accuracy in obtaining results for all surface water sampling points across England. The names and BNG coordinates of these sampling points can be obtained from the publicly available EA archive database.

12. Key Findings, Strength and Weakness

12.1. Key Findings

- The EA water quality monitoring dataset includes 70 million observations (2000-2024) from 55 rivers and 25 aquifers with 3,261 physico-chemical determinands and spatiotemporal metadata.
- Phosphate and orthophosphate account for only 0.2% and 2% of those observations, respectively.
- Clustering: Identified 6 chemically distinct groups among the determinands using K-Means++ and agglomerative clustering.
- Advanced techniques: Using SMILES text, Morgan Fingerprints, ChemBERTa (chemical-specific transformer), ChemDataExtractor, and PubChemPy enabled clustering of 370+ orthophosphate related determinands.
- Dimensionality reduction: PCA retained 93% variance after converting SMILES to Morgan Fingerprints (molecular bitmaps). To achieve 93% explained variance, the optimal number of components is 356 (refer to Figure 8). Essentially, the Morgan fingerprints representing the molecular structure of a chemical, obtained using RDKit, form a 2168 by 2048 matrix (refer to section 4.2 Feature-Selection Methodology). This means the structure is represented for a chemical as a 1 by 2048 matrix in computer-readable bit form, which is a critical step. However, we do not need such a large feature matrix that shall go into the K-Means++ to draw clustering. So PCA was employed. PCA reduces it to 356 components for each chemical, allowing 93% of the variance to be understood, compared to 100% using the original 2048 components.
- **Hybrid feature selection:** A comprehensive hybrid approach was employed, which included feedback from domain experts. We utilised advanced feature

selection methods such as agglomerative clustering and traditional correlation-based analysis. These methods, combined with multiple experiments, resulted in finalising representative features.

- Model performance: LightGBM model outperformed XGBoost, Random Forest, ANN and liner regression models achieving:
    - MSE 0.00123 mg/l
    - MAE 0.01 mg/l
    - NRMSE 8%
- Validation & Explainability:
    - Robust Splitting: spatiotemporal splits prevented data leakage.
    - SHAP analysis: The objective of the SHAP implementation is to give insights into how the model determines the orthophosphate value using the weights of the supplied input. Using this, a domain expert may be able to make a decision whether to trust specific prediction or disregard. Example insight: Ammonia contributed 76% of prediction variance.

12.2. Strengths

- Predicts orthophosphate at all sampling points, including the 72%[38] sample points that are currently unmonitored for orthophosphate (MAE = 0.02, NRMSE = 7%)
- The model demonstrates the ability to predict orthophosphate levels by utilising 45 distinct features, trained from data collected at 18,000 sampling locations. It achieves predictions an NRMSE of less than or equal to 8%. Notably, the model retains its performance within an 8% NRMSE, even when the number of features is reduced to just 11 (refer to 8c-2) during validation at a single sampling location. It provides predictions even when data is limited. These performance metrics should be interpreted in the context of the above (section 2) and below (section 13.1 and 16.4) discussions about Limit of Detection values and our cap of 0.5 mg/L. .
- Combines ML with domain expertise for actionable insights
- Transparent via SHAP, critical for regulatory compliance

13. Limitations and opportunities

13.1. Orthophosphate model

The model predicts orthophosphate values in mg/l (determinand Notation: 0180, Description: *orthophosphate, reactive as P*), not total phosphorus (Notation 0348,

38 The training set for surface samples we used has samples from around 18,000 sampling points. Using this we calculated 72% as unmonitored sampling points out of the total 64,000 sampling points been monitored in England

Phosphorus, Total as P) which includes the portion bound to sediments. This distinction is essential because orthophosphate measures the reactive phosphorus available for immediate uptake by organisms, whereas total phosphorus includes all forms, including those that are not readily available. It is important to understand this difference between orthophosphate and total phosphorus, as it can affect the interpretation of water quality data and subsequent environmental management decisions.

The model performance is depending on being supplied an existing EA Sampling point (from WIMS dataset). If not supplied, the model will still run and produce an estimation, but it will do so without any location specific knowledge, and the performance is likely to be affected. Suggestions for incorporating new sites is discussed in section 11.

To enhance real-world applicability, we recommend considering incorporating the "Limit of Detection (LOD)" as a feature. The LOD refers to the lowest concentration of a substance that can be reliably detected but not necessarily quantified. Integrating LOD can help the model distinguish between actual zero measurements and values that were limited by the measurement equipment and methods.

Another limitation of our model is the initial cap selection of 0.5 mg/L, based on diverse sample types. Although we intended to apply the models only to surface samples from various points across England, consultation with a domain expert suggests that this cap may need revision. As part of our validation (section 16.4) we did trial the effects of revising this cap, but more work is still needed to understand the full potential of this.

Furthermore, the model currently assigns a value of zero to both 1) specific data points in the time series where data is missing and 2) all data points for determinands that were not sampled. Whilst it is assumed that item 2 is likely to have negligible impact of model predictions, further work is needed to assess the impacts of item 1 on the contribution of genuine measurements of zero (such as for temperature) to model performance and also how it might affect the weightings assigned to measurements that are very near to zero.

The model supplements, rather than replaces, water quality monitoring. By using advanced algorithms and data analytics, it provides additional insights and predictions about water quality trends and potential issues, which helps in making informed decisions and implementing timely interventions. Traditional monitoring methods are still essential for collecting real-time data and validating the model's predictions.

13.2. EA's WQ monitoring data limitations

13.2.1. Limited Coverage:

Sampling networks may not represent all river types, especially smaller rivers and streams. This could lead to inadequate representation of the catchment patterns and lower model performance in those areas compared to where there is adequate representation.

13.2.2. Data Gaps:

Missing spot samples due to disease outbreaks or lockdowns create data gaps that may significantly affect the model's performance.

14. References

- Chithrananda S, Grand G, & Ramsundar B. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. Huggingface link
- Ke G, et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS
- Ribeiro, M. T. et al. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938
- Swain, M. C., & Cole, J. M. "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature", J. Chem. Inf. Model. 2016, 56 (10), pp 1894–1904 10.1021/acs.jcim.6b00207
- PubChemPy global chemical database reference
- K-Means ++ : Advantage of careful seeding

15. Annexes

15.1. Train – Test – Validation Datasets:

In machine learning, a model's performance is evaluated using three datasets: training, test, and validation sets. The training set is used to train the model and allow it to learn patterns from the data. The test set helps in fine-tuning the model's hyperparameters and assessing its performance during training, preventing overfitting. The validation set provides an unbiased evaluation of the model's final performance and generalization ability on unseen data.

Training Set:

This is the primary dataset used for teaching the model how to make predictions. The model learns from the examples in this set and adjusts its parameters to minimize errors.

Test Set:

The test set is used to evaluate the model's performance after it has been trained. It helps in choosing the best model and optimizing its hyperparameters (settings that control the learning process). By comparing the model's performance on the test set, we can adjust parameters to prevent overfitting[39].

Validation Set:

This set is completely independent of the training and test sets. It's used as a final assessment of the model's performance on unseen data to determine how well it

39 Overfitting results in model learning the training data well but performs poorly on new data

generalizes to real-world scenarios. Generally carried out in collaboration with domain experts, the validation set is crucial for determining the model's final accuracy and is not used during the training or hyperparameter tuning process.

In essence, Training: The model "learns" from the data. Test: The model is "tuned" and "evaluated" to ensure it's not overfitting. Validation: The model's final "performance" is assessed on new, unseen data

15.2. SHAP:

SHapley **A**dditive ex**P**lanations, is a method for explaining the output of machine learning models by assigning importance scores to each input feature. The Beeswarm plot of SHAP is designed to display an information-dense summary of how the top features in a dataset impact the model's output. Each instance the given explanation is represented by a single dot on each feature row. The x position of the dot is determined by the SHAP value.

16. Model Validation

16.1 Primary purpose of the validation:

The overarching purpose of this validation exercise is to assess the performance of the Open Orthophosphate Model using previously unseen water quality data. By exposing the model to entirely new datasets, this process aims to provide a realistic evaluation of its predictive accuracy and reliability under real-world conditions—beyond the bounds of its original training environment.

The validation is structured around two principal data sources: comprehensive monitoring data from the Environment Agency (EA), and supplementary citizen science datasets (collecting orthophosphate measurements) and continuous EA sonde data (collecting proxy parameters). By incorporating these independent sources, the assessment is designed to cover a broad range of scenarios, catchment types, and water quality conditions that the model may encounter in operational use.

Ultimately, this approach allows for an in-depth understanding of the model's strengths, its potential limitations, and the circumstances under which its predictions remain robust or may require caution. The findings from this validation will guide future refinements, ensuring the model delivers accurate orthophosphate predictions that are fit for purpose across diverse environmental monitoring applications.

16.2 Validation using Environment Agency's Water Quality data

We have ingested one year of data (01 June 2024 to 17 July 2025) from the Environment Agency's (EA) water quality archive, and these are data the model has not seen before (i.e.,

has not been trained on). This provides one full year of data to examine model strengths and weaknesses, compared to the training accuracy.

Pre-requisite for data: The model utilises 45 determinands to predict orthophosphate values. It requires input for each determinand, using either the actual measured values or zero where data are unavailable. Ideally, providing actual values for at least 25 of these determinands will improve prediction reliability. Key parameters necessary for accurate orthophosphate prediction include all six Section-82 parameters as well as others such as Temp Water, Cond@ 25C, Ammonia (N), N-Oxidised, Nitrate-N, Nitrite-N, Alky pH 4.5, and BOD ATU[40]. For additional required determinands and the training accuracy details, please refer to the Table 9 and '8c-1' sections of this report.

Validation categories & performances: Validations for the following 9 different specific categories of catchments have been executed, based on one year of data from the EA's water quality monitoring database (WIMS) (216 sampling points across England in total; sampling points for each case are shown in brackets). For more detail see table 14.

All available catchments:

1. One year of EA water quality monitoring data (216 sampling points)

Sampling point data for varied categories:

2. High Base Flow Index (BFI) (31 sampling points).
3. Low BFI (33 sampling points).
4. A UK benchmark catchment[41] (16 sampling points).
5. A catchment not a UK benchmarking catchment (20 sampling points).
6. A rural catchment (23 sampling points).
7. An urban catchment (17 sampling points).
8. A dry catchment (20 sampling points).
9. A wet catchment (20 sampling points).

WIMS Dataset: Data points for measurements taken between 01 June 2024 and 17 July 2025 were considered. These data points were obtained from the Environment Agency's data archive (Located under EA's data services platform). The data set was filtered to have the measurement that pertains to only 'River / Running Surface Water'.

[40] All parameter labels are from the EA water quality archive column "determinand.label".

[41] Benchmark catchments can be considered reasonably free from human disturbances such as urbanisation, river engineering, and water abstractions, so are 'near natural'. https://nrfa.ceh.ac.uk/hydrometry-uk/benchmark-network

Table 14 summarises the methodology for selecting Environment Agency (EA) monitoring points for orthophosphate data (validation 1 and 2). This methodology included targeted selection of EA monitoring points based on the number of orthophosphate measurements recorded, ensuring robust data coverage for both high and low BFI operational catchments as well as UK benchmark catchments. This stratified approach allowed for a comprehensive analysis across varying hydrological conditions, thereby enhancing the predictive accuracy of the orthophosphate model for diverse sampling point locations.

Component	Approach
EA monitoring points	• Ingested all records of orthophosphate (Determinand notation: 180) from the EA water quality database for 'RIVER / RUNNING SURFACE WATER' collected between 01 June 2024 and 17 June 2025. Calculated the frequency of orthophosphate measurements for each unique monitoring point ID. Generated a shapefile and visualized the data in GIS, with points color-coded by the number of orthophosphate measurements. Ingestion was focused on monitoring points with more than 40 orthophosphate measurements.
High BFI[42]	• WFD[43] operational catchments were grouped based on the BFI of NRFA[44] gauging points within them and mapped in GIS by BFI category. EA monitoring points with over 40 orthophosphate measurements in high BFI areas (0.636–0.955) were selected.
Low BFI	• WFD operational catchments were grouped by BFI based on NRFA gauging points, then mapped in GIS by BFI category. EA monitoring points with over 40 orthophosphate measurements in low BFI areas (0.194-0.388) were selected.
UK Benchmark Catchments	• Plotted benchmark catchments (NRFA subset) and CAMELS-GB boundaries. Selected EA monitoring points with more than 40 orthophosphate measurements within each benchmark[45] (CAMELS-GB) catchment draining into the benchmark point.
Non-UK Benchmark Catchments	• Plotted benchmark catchments (NRFA subset) and CAMELS-GB boundaries. Selected EA monitoring points with more than 40 orthophosphate measures outside of benchmark catchments (where CAMELS-GB catchment drains into benchmark point).

[42] Base Flow Index
[43] Water Framework Directive
[44] National River Flow Archive
[45] Benchmark catchments can be considered reasonably free from human disturbances such as urbanisation, river engineering, and water abstractions, so are 'near natural'. https://nrfa.ceh.ac.uk/hydrometry-uk/benchmark-network

Wet catchment	• Mapped Met Office 1-km annual average rainfall (1991–2020) and selected EA monitoring sites with over 40 orthophosphate measurements in England’s wettest regions.
Dry catchment	• Mapped Met Office 1-km annual average rainfall (1991–2020) and selected EA monitoring sites with over 40 orthophosphate measurements in England's driest areas.
Urban	• Mapped the OS built-up areas in GIS together with the WFD river water body catchments, and visually identified EA monitoring points that have more than 40 orthophosphate measurements and significant built-up areas within their upstream catchments.
Rural	• Mapped the Ordnance Survey built-up areas in GIS alongside the WFD river water body catchments. Visually identified Environment Agency monitoring locations with more than 40 orthophosphate measurements, where the upstream catchment contains little to no built-up area.

Table 14. methodology for selecting Environment Agency (EA) monitoring points

Validation 1: One year of EA water quality monitoring data (Data that were not used in training the orthophosphate model).

Measured orthophosphate range: 0.001-0.5 mg/l	Location Categories	Sampling point Locations (Total)	MSE	MAE	RMSE	Total predictions
Performance metrics stay on par with the training metrics	Whole 1 year	216	0.00257	0.03	0.05	10628
	Data selection: Orthophosphate above 0 & below 0.5 mg/l | all metrics are in mg/L					

Table 13: Performance metrics using the EA's one-year WQ dataset (multi sites)

Observation: The number of the predictions exceeded 10,000. The RMSE is 0.05 mg/l, similar to the training prediction. The scatter plot below (Figure 16) illustrates this.

This shows that the data contain many small values, with a standard deviation indicating some variability. An RMSE error of 0.05 might seem minor, but it represents a 100% error if the actual value is 0.05. Thus, comparing the error to the data scale becomes very essential in this context.

The Mean Absolute Error (MAE) is 0.03, meaning the model's predictions deviate by 0.03 units on average from the true value. The average actual value is 0.049, making the error about 60% of the average measurement. A better comparison is to the median (0.025) because in this case, the MAE is only slightly larger than the median itself. However, including other metrics will provide a fuller picture, since MAE can be skewed by a few large errors and does not capture the model performance on its own.

The Root Mean Square Error (RMSE) is 0.05, indicating that a typical error is about 0.05 units. RMSE measures average error but gives more weight to larger errors (it penalises big mistakes more severely). This is the most important metric in our analysis and is identical to the mean value (0.049) and smaller than the standard deviation (0.063). The Mean Squared Error (MSE) is 0.00257, confirming that the squared errors are small. It is primarily used to calculate the RMSE.

The Normalised RMSE (NRMSE) is 10.15%, indicating the average error is about 10% of the overall average readings if calculated using the following formula:

$$NRMSE = \frac{RMSE}{(Observed\ Orthophosphate\ range)} * 100$$

Calculating the NRMSE with the mean value (0.049) instead of the max-min (range) gives an NRMSE of 102%. Meaning that the average errors is about 102% of the mean orthophosphate readings.

$$NRMSE = \frac{RMSE}{(Mean\ Observed\ Orthophosphate)} * 100$$

Scatter plot visualisation:

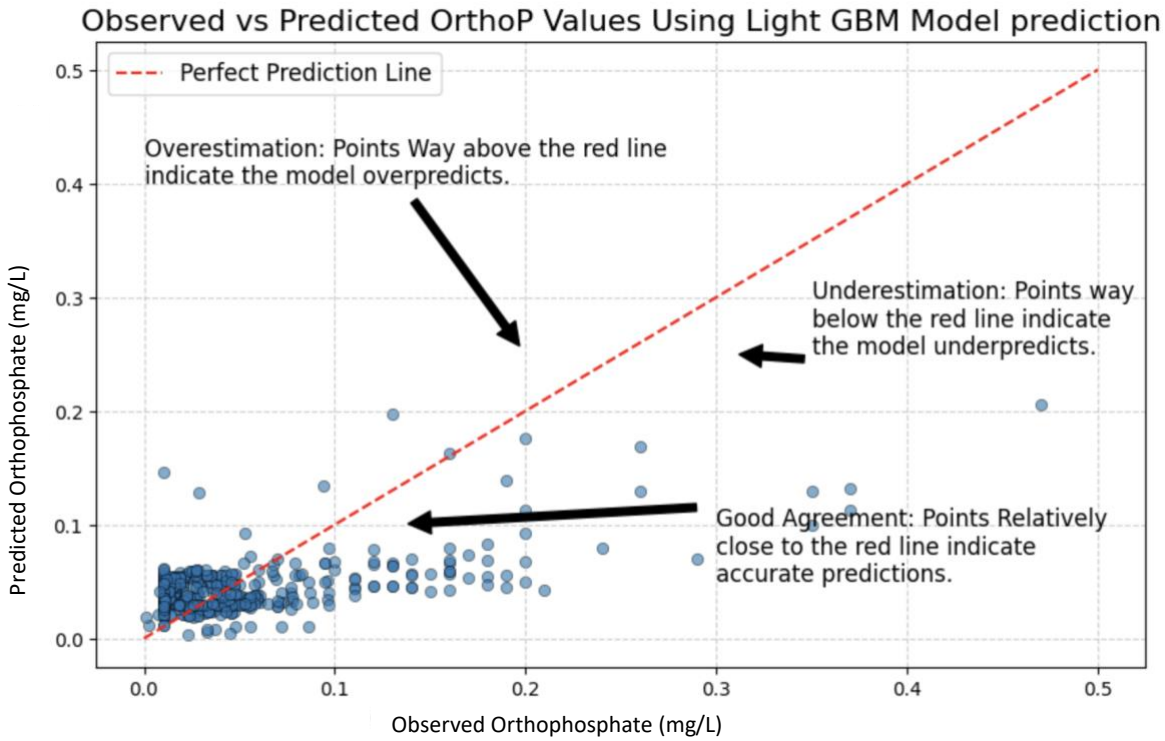


Figure 16: Orthophosphate measures: Actual vs Predicted (multiple sites)

Validation 2: Sampling point data for varied categories

Below we highlight the model performance for the following catchment types.

1. High BFI Vs Low BFI catchments
2. Benchmark vs Non benchmark catchments
3. Rural vs Urban catchments
4. Dry vs Wet catchments

Inference: The model's performance on the validation data generally mirrors its results during training (see figure 17 to 24). It shows consistent results in predicting orthophosphate concentrations under normal conditions, indicating that it can apply learned patterns to unseen datasets. However, its predictive accuracy declines at sites experiencing extreme events—such as exceptionally high or low orthophosphate readings. This indicates that the model's assumptions and training data do not adequately capture the full range of variability present in these outlier scenarios. Therefore, while the model performs consistently for standard situations, its predictions should be treated with caution when applied to data from sites known to be affected by unusual or extreme events, where its ability to generalise is restricted. Below are charts showing the details of the performance metrics for the different catchment types outlined above.

Operational Catchment ID	EA monitoring points with >= 40 measures of ortho-P
3527	AN-WEN180
	AN-WEN200
	AN-TUD070
	AN-WEN040
	AN-WEN010
	AN-WEN020
3331	AN-BRN005
3249	AN-37M28
3064	AN-27M43
3282	SO-E0001389
	SO-E0007230
3121	SO-E0000135
	SO-E0000139
3533	SO-F0002997
	SO-F0003014
3536	TH-PWER0104
3261	TH-PLDR0054
3242	TH-PKER0116
	TH-PKER0014
3216	SW-50250102
	SW-50280585
	SW-50240116
	SW-50210209
	SW-50220110
	SW-50220136
	SW-50230145
3140	SW-50350235
	SW-50350521
	SW-50350414
	SW-50360430
	SW-50331119

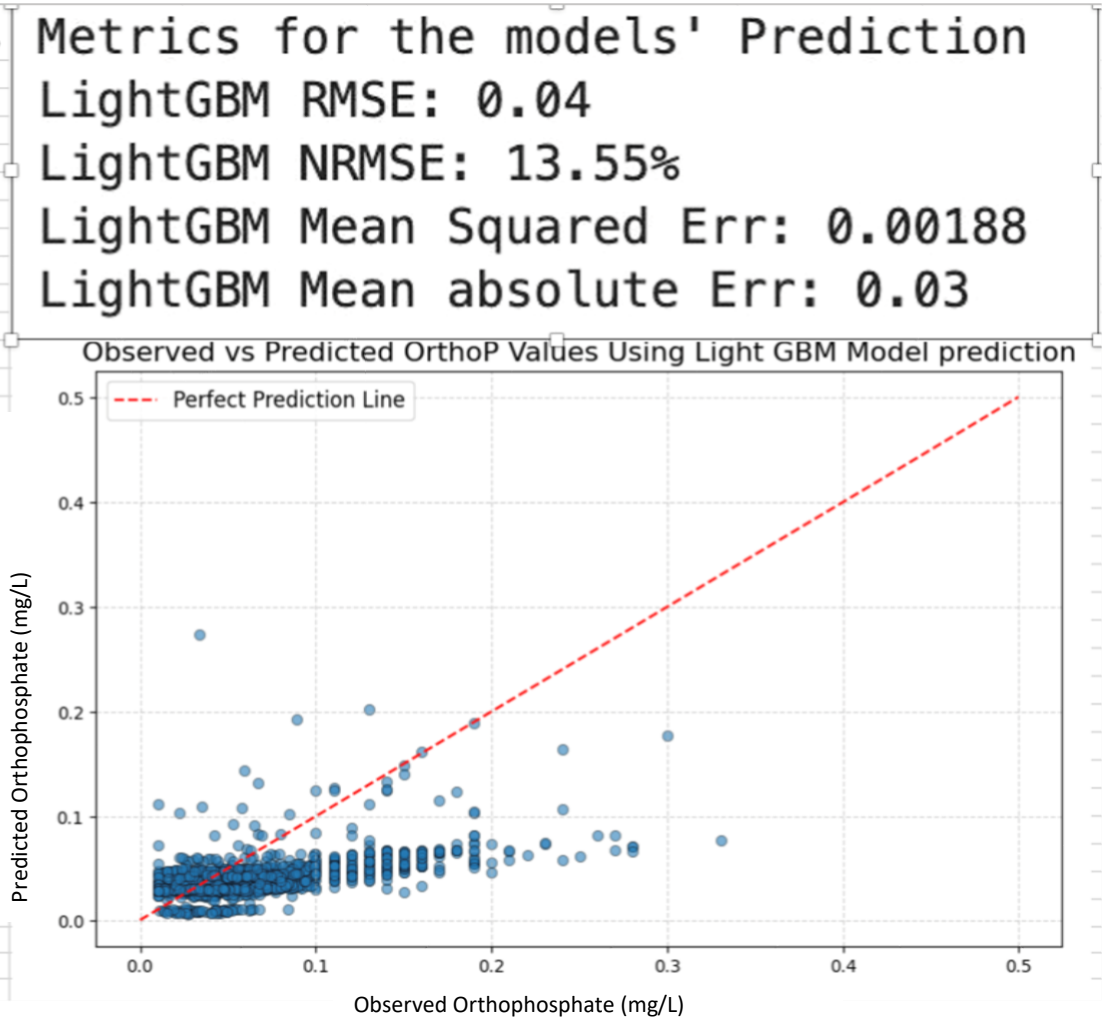


Figure 17: High BFI sampling points.

Operational Catchment ID	EA monitoring points with >= 40 measures of ortho-P
3145	NW-88004949
3178	NW-88006458
	NW-88006481
	NW-88006480
3515	NE-42500138
	NE-42500182
3334	NE-43100177
	NE-43100099
3372	NE-43100153
3107	NE-42300181
3505	NE-49100650
	NE-49100136
	NE-49100031
3507	NE-49700078
	NE-49705096
3316	NW-88003418
	NW-88003417
3223	NW-88003452
	NW-88003449
3051	NW-88003843
	NW-88003867
	NW-88003840
	NW-88003872
3018	MD-07716050
3483	SO-F0002631
3031	SO-E0000553
	SO-E0017059
	SO-E0017060
3275	TH-PLER0157
	TH-PLER0131
3077	AN-TE0130
	AN-TE0183
	AN-CH1080

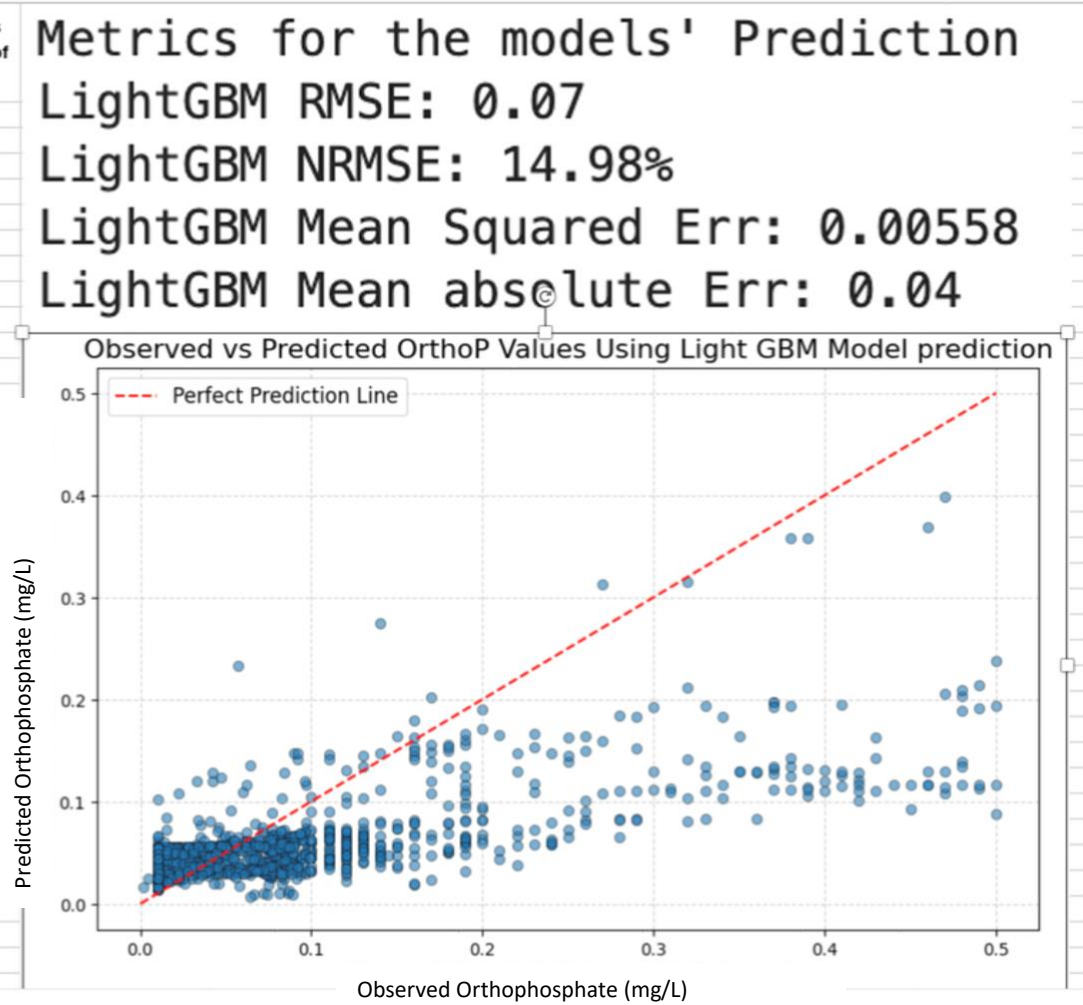


Figure 18: Low BFI sampling points

Camels_GBl Catchment ID	EA monitoring points with >= 40 measures of ortho-P
34011	AN-WEN040
	AN-WEN010
	AN-WEN020
40005	SO-E0017060
	SO-E0047059
	SO-E0000553
53018	SW-Z1520101
53006	SW-Z2050102
22001	NE-42300085
74001	NW-88004949
32003	AN-HARP140B
41003	SO-F0002099
46003	SW-70723535
	SW-70724308
	SW-70720287
	SW-70725509

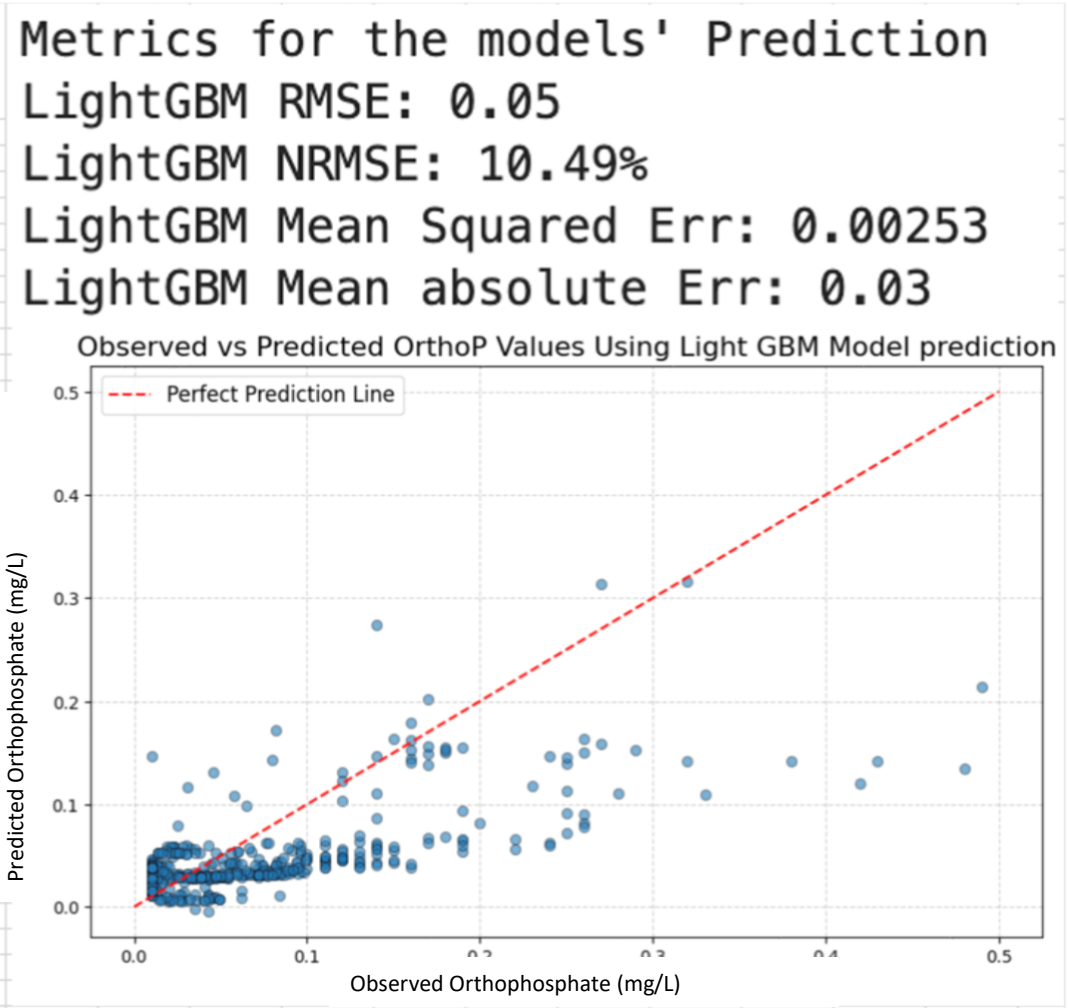


Figure 19: UK Benchmark catchments

Camels_GBl Catchment ID	EA monitoring points with >= 40 measures of ortho-P
48005	SW-81942260
49001	SW-82521235
	SW-82530105
	SW-82528159
42004	SOG0006183
	SO-G0003918
	SO-G0004084
40018	SO-SSN0715
	SO-E0000135
34004	AN-WEN200
33014	AN-37M28
28024	MD-47519900
	MD-48189390
21009	NE-41000098
	NE-41000038
	NE-41000372
	NE-41000051
22009	NE-42300181
23008	NE-43100153
76007	NW-88006288

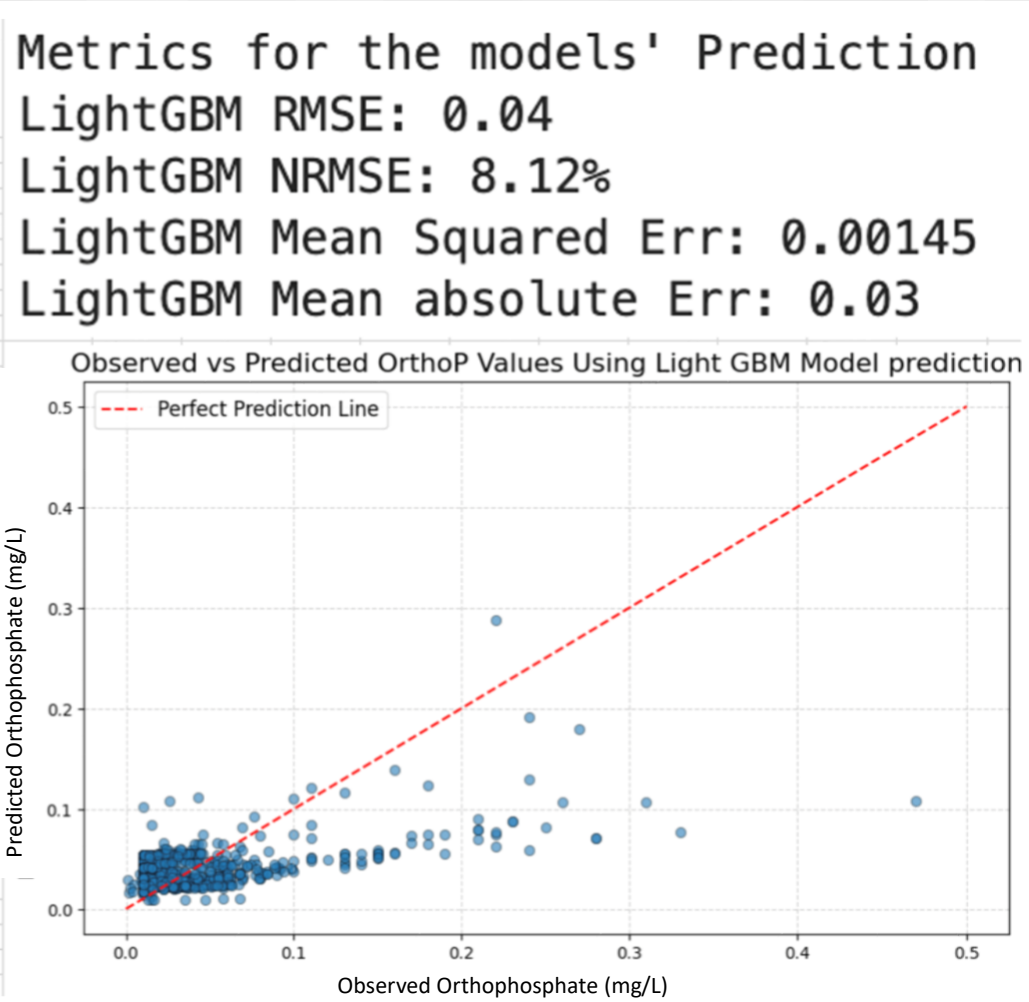


Figure 20: Non-UK Benchmark catchments

EA monitoring points with >= 40 measures of ortho-P
SW-82010156
SW-81940830
SW-81942260
SW-81522004
SW-82530105
SW-82528159
SW-91241720
AN-BR0316
AN-TE0183
MD-47519900
NE-49600482
NE-49000476
NE-49000127
NE-49000488
NE-42400046
NE-42500138
NE-42500182
NE-43100177
NE-43100153
NW-88023146
NW-88005567
NW-88004949
NW-88005121

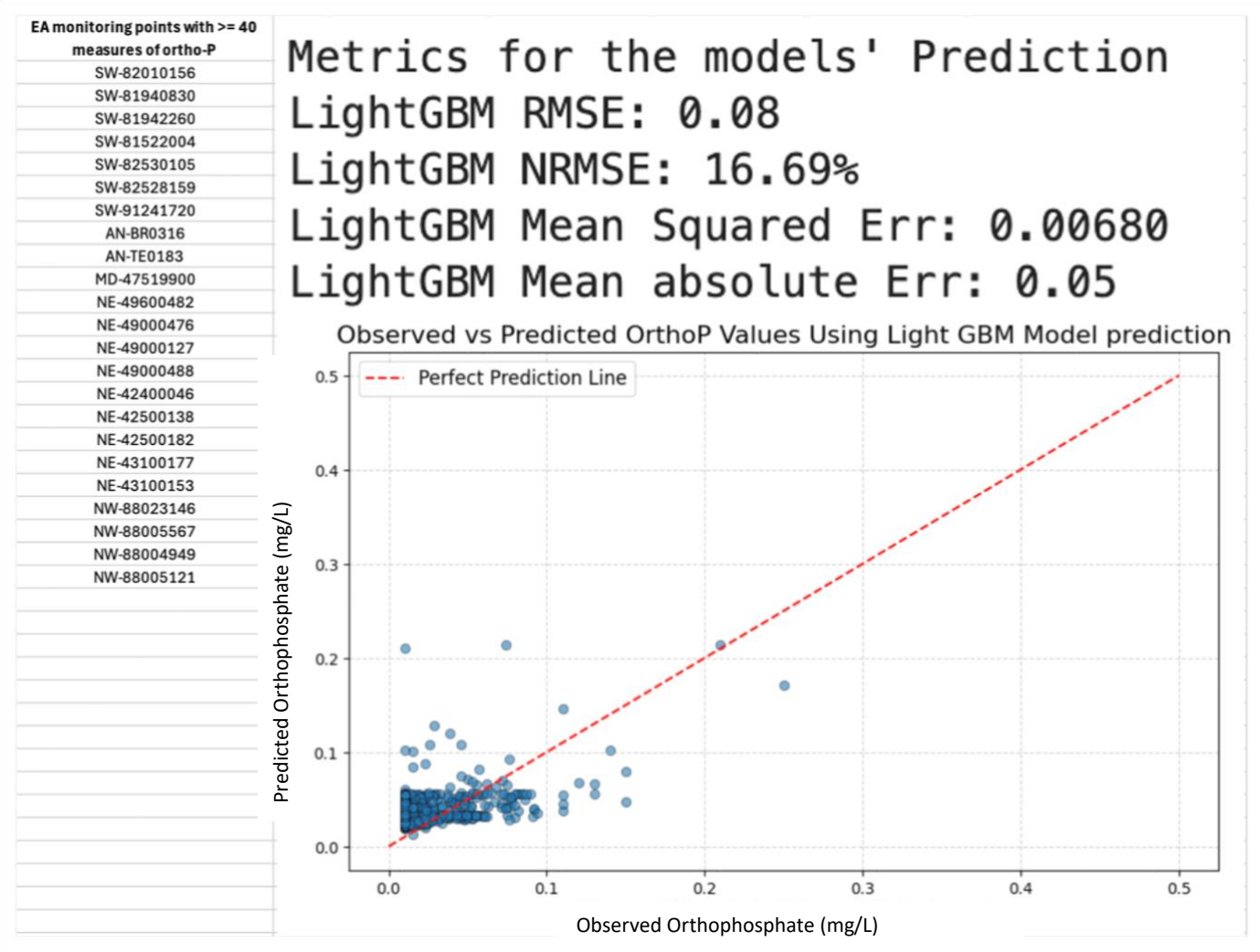


Figure 21: Rural sampling points

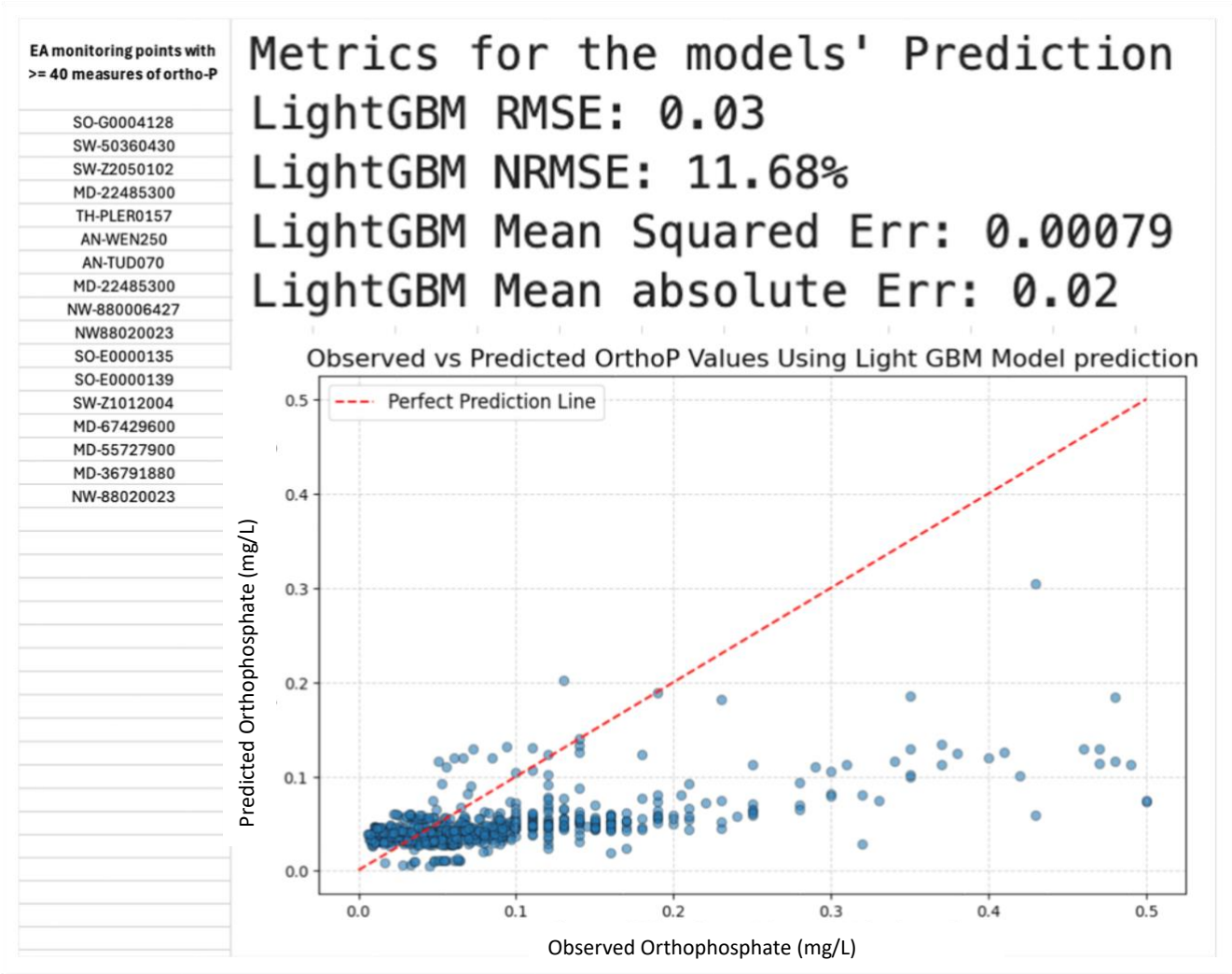


Figure 22: Urban sampling points

EA monitoring points with >= 40 measures of ortho-P
AN-OAE047
AN-DEB020
AN-ALD020
AN-ORE040
AN-WEN200
AN-TUD070
AN-WEN250
AN-BRN005
AN-WEN040
AN-WEN010
AN-WEN020
AN-37M28
SO-E0000553
SO-E0000616
SO-E0017060
SO-F0002067
SO-F0002111
SO-F0002332
SO-F0002290
SO-E0002374

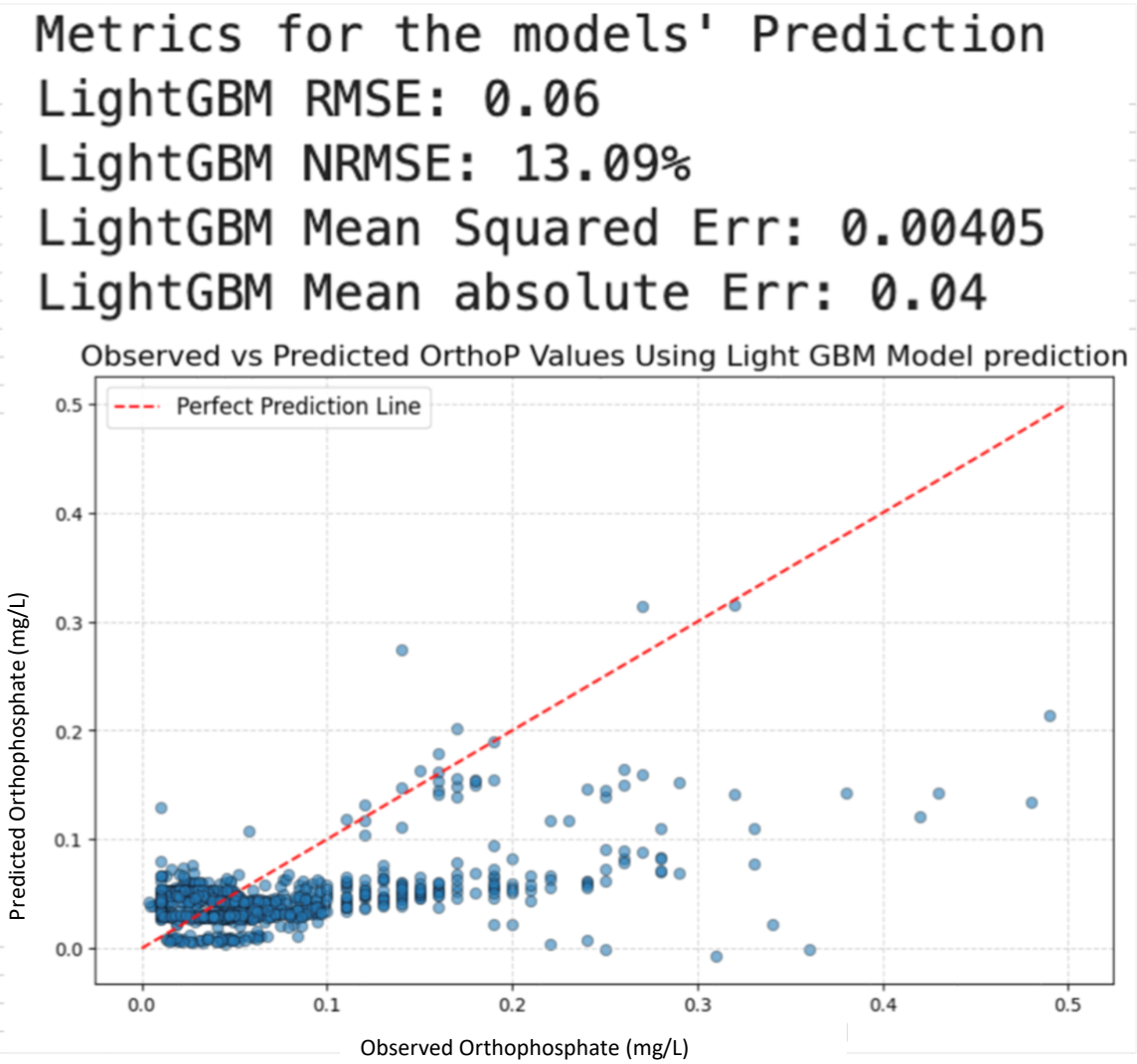


Figure 23: Dry catchments

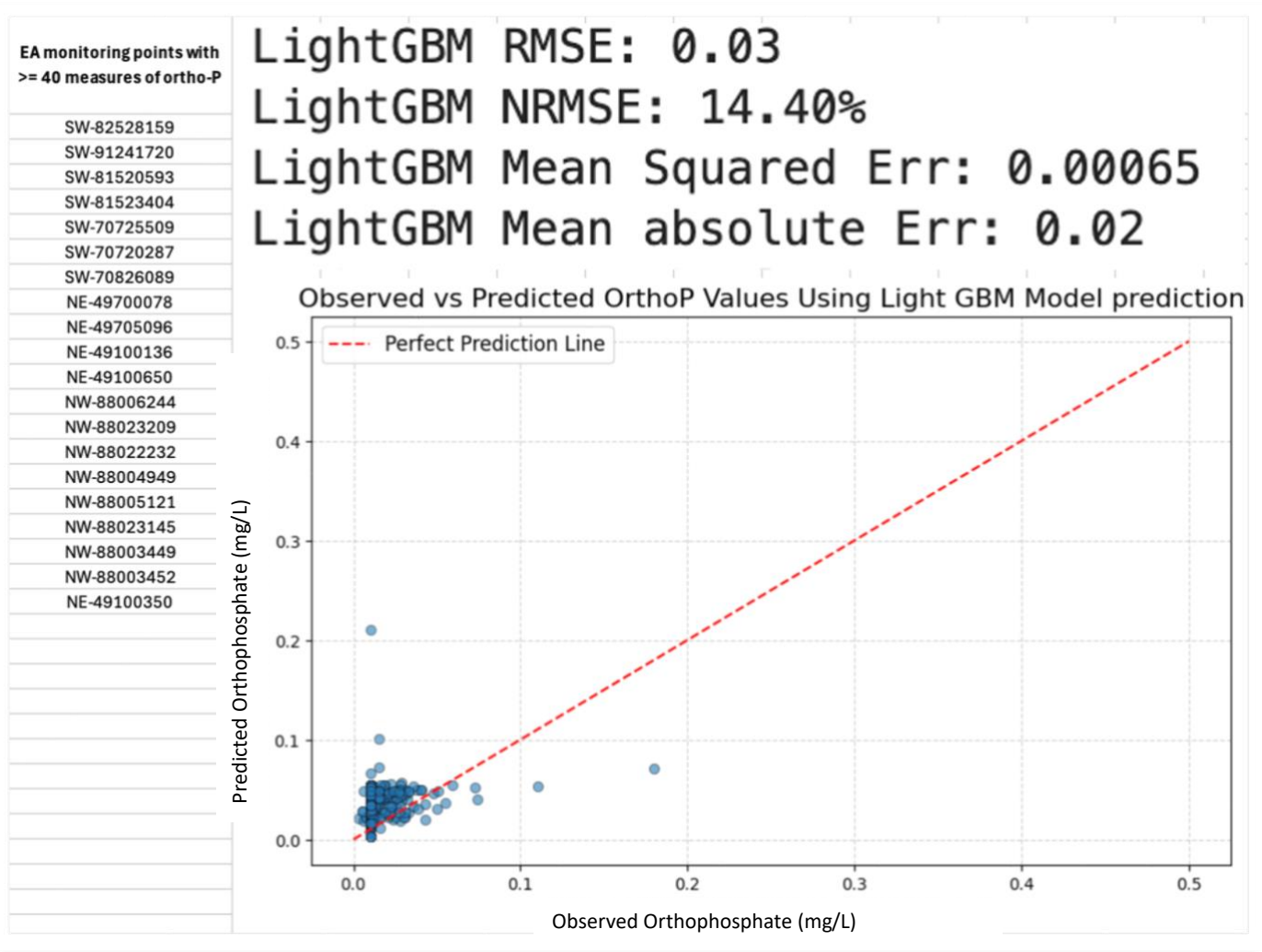


Figure 24: Wet catchments

Catchment type	RMSE	NRMSE	Mean Square Err.	Mean Abs. Err.
All catchments	0.05	10.15%	0.002	0.03
High BFI	0.04	13.55%	0.002	0.03
Low BFI	0.07	14.98%	0.006	0.04
Benchmark	0.05	10.49%	0.003	0.03
Non-Benchmark	0.04	8.12%	0.001	0.03
Rural	0.08	16.59%	0.006	0.05
Urban	0.03	11.68%	0.001	0.02
Dry	0.06	13.09%	0.004	0.04
Wet	0.03	14.40%	0.001	0.02

Table 15: Overview of validation results (validation 1 and 2).

Table 15 outlines the validation of the model across catchment types. The overall performance is moderate (RMSE ≈ 0.05; NRMSE ≈ 10.2%). The model performs best for non-benchmark sites (NRMSE 8.1%, low MSE 0.001) and for urban and wet catchments (RMSE 0.03; MAE 0.02), indicating it generalises well to impacted or hydrologically buffered settings. Errors increase for rural (worst: RMSE 0.08; NRMSE 16.6%; MAE 0.05) and low-BFI catchments (RMSE 0.07; NRMSE 15.0%). High-BFI and benchmark groups sit near the overall average (RMSE 0.04–0.05).

In short, the model is more reliable for non-benchmark/urban/wet catchments and struggles most in rural and low-BFI systems.

16.3 Validation using citizen science data

Further validation has been undertaken using EA sonde data (NH_4-N, DO, Water Temperature, pH) to infer orthophosphate, which has been validated against citizen science data collected by the RDMAI consortium in the same or nearby points. This has been done to see how well the model performs when provided with out-of-sample data from a third-party provider, in this case, the EA sonde data. This approach has several limitations and dependencies that we will outline below.

Data conversion and mapping

The data retrieved from EA Sondes was mapped to determinand notations from the EA WIMS dataset to match the input data needed to run the model (table 16). This mapping was done in the following way:

EA Sonde data	WIMS determinand notation	WIMS determinand name
Ammonium as N (NH4+ as N)	0111*	Ammoniacal Nitrogen as N
	0119*	Ammonia un-ionised as N
Dissolved Oxygen (mg/L)	9924	Oxygen, dissolved as O2
Dissolved Oxygen (%)	9901	Oxygen, dissolved, % Saturation
Water temperature	0076	Temperature of Water
pH	0061	pH

Table 16: Overview of mapping between EA Sonde and EA WIMS parameters.

*Ammonium as N (NH4+) collected by the EA sonde was used to estimate two determinands useful as model input parameters. This was done using the following conversion formulas:

$$Ammonia\ unionised\ as\ N = \frac{NH4^{+}\ as\ N}{10^{pKa-pH}}$$

$$pKa = 0.0901821 + \frac{2729{,}93}{Temperature + 273{,}15}$$

$$Ammoniacal\ Nitrogen\ as\ N = NH4^{+} as\ N + Ammonia\ unionised\ as\ N$$

These conversions and approach were agreed in close collaboration with project partners. It is important to note that converting parameters from one source into different parameters useful for the model introduces some uncertainty. The supply of only 6 out of 45 determinands also introduces some inherent limitations. However, for this case, we are willing to introduce this uncertainty to see how the model handles sparse data collected from different sources. Furthermore, estimating orthophosphate with a limited set of proxy parameters that can realistically be measured has been the overarching goal for this building block.

Each EA Sonde location was mapped to the nearest EA WIMS sampling point as the Open Orthophosphate Model needs an existing EA WIMS sampling point to run (see section 13).

Using the above determinands, the model was run to estimate Orthophosphate in the given locations. These estimates were then validated against Orthophosphate measurements collected by citizen scientists in the same location as the EA sonde. Below, we provide an overview of the results from this validation step.

Results of validation – EA Sonde and Citizen science data

In this validation step, we ran the model in 6 locations with citizen science samples: Pix Brook (59 samples), River Severn (5 samples), River Teme (3 samples), Warren Dike (6 samples), River Rede (1 sample) and South Delph (7 samples).

To make the predictions (orthophosphate as P) comparable with the citizen science samples, each predicted value was multiplied by 3.06[46] to account for the molecular weight-difference between Orthophosphate and Orthophosphate as P.

These values were then compared with the citizen science samples to calculate MAE, RMSE and R^2 for each individual sampling point and combined (see table 17).

[46] https://www.cheminc.com/post/understanding-phosphorus

	Mean OrthoP (from Citizen science)	**Mean Absolute Error** (mg/L)	R^2	**RMSE**
All sampling points (81)	1.79 mg/L	1.72	0.025	2.32
Pix Brook (59)	2.3 mg/L	2.2	0.0007	2.69
River Severn (5)	0.078 mg/L	0.073	0.599	0.08
River Teme (3)	0.013 mg/L	0.052	1	0.05
Warren Dike (6)	0.166 mg/L	0.111	0.301	0.12
River Rede (1)	0 mg/L	0.0008	-	-
South Delph (7)	1.19 mg/L	1.054	0.054	1.06

Table 17: Results from citizen science / EA sonde validation

The results indicate that the model struggles when orthophosphate measurements are high (in particular Pix Brook), whereas the MAE and R^2 improves when the orthophosphate is lower. This validation exercise highlights some limitations to the model when estimating orthophosphate using data from EA sondes instead of WIMS data. Therefore, the results of the model when used in this context should be viewed with caution.

The RDMAI Open Orthophosphate model was trained using EA water quality data. These data have a very low temporal resolution – typically 1 to 2 readings per month. Therefore, it is likely that the EA monitoring data miss spikes in orthophosphate associated with for example a rainfall event. The sonde data have a much higher temporal resolution. Therefore, the sonde data will capture these spikes in pollution associated with ephemeral events. This points to an inherent limitation of training models to the EA data.

16.4 Test of revising cap

In all validation exercises (including WIMS and out-of-sample), it was apparent that the model struggles more when the actual orthophosphate is high. We believe that this might be due to the cap previously implemented, where training data used had an orthophosphate concentration within 0.5 mg/l.

To test this assumption, we re-trained the base model with a new orthophosphate concentration cap on training data of 10 mg/l. A summary of the performance of the re-trained model is given below. These preliminary results indicate that the base model performs better at the original cap. However, this does not rule out possible benefits of revising this cap, and the implementation of the new cap might be a better reflection of the model performance when deployed in the real world. It is likely that the initial cap was unsuitable

for model development. More experimentation and analysis is needed to understand how to improve the model performance in instances of high orthophosphate levels.

As mentioned in section 13.1, we suggest removing the current 0.5 mg/L cap and removing (or flattening) the artificial spike of data in the lower ranges caused by the limit of detection (LOD) (see section 2). This might improve the model performance and give a better understanding of what performance to expect when deploying the model in the real world.

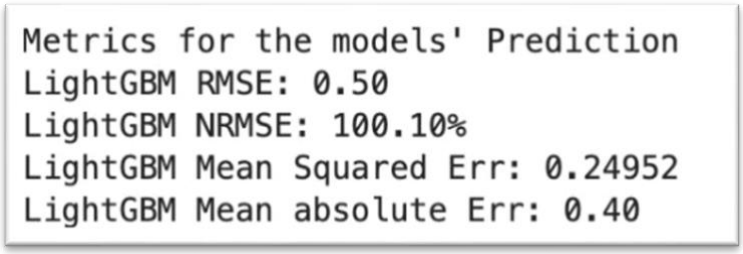


Figure 25: Validation of Model Re-Training performance using a revised orthophosphate cap of 10 mg/l

17. Glossary

AI/ML: Artificial Intelligence and Machine Learning

ACF: Autocorrelation Function & **PACF:** Partial Autocorrelation Function are functions identifying patterns, trends and order of autoregressive models

Adjusted R-Squared: Adjusted R-Squared is a dependable metric for performance evaluation, as it provides insights into the number of independent variables that offer substantive information

ANN: Artificial Neural Network

EA: Environment Agency, a government agency responsible for environmental protection and regulation in England

LLM: Large Language Model

LOD: Limit of Detection

MAE: Mean Absolute Error

Molecular similarity analysis: Molecules with similar structures often exhibit predictable interactions. These interactions may involve binding to identical biological receptors, undergoing analogous metabolic degradation processes, coexisting within environmental

matrices such as aquatic systems, and either competing or cooperating within microbial communities or ecological networks

MSE: Mean Squared Error. This metric measures the average squared difference between predicted and actual values. For instance, extremely low value (0.000123) indicates that the model is making accurate predictions, with the errors being exceedingly small. Squared errors penalise larger errors more heavily, so a low MSE suggests the model is making accurate predictions, even on instances with larger errors.

NLP: Natural Language Process

NRMSE: Normalized Root Mean Square Error

OrthoP: Orthophosphate

P: Phosphate

PCA: Principal Component Analysis

Q metrics: Also known as Q-statistics or Q-values, are a set of metrics used to assess the explained variance, bias, and variance ratio in model predictions. These metrics help in understanding how well a model generalizes and whether it suffers from underfitting (high bias) or overfitting (high variance). Explained variance measures how much variance in the target variable is explained by the model. Bias measures the average difference between predicted and actual values. Variance ratio compares unexplained variance to explained variance

R2: R-Squared is a performance evaluation metric that measures how well regression model captures variability in data

RMSE: Root Mean Square Error

Scikit-Learn mutual_info_regression for feature selection: Mutual information measures the dependency between variables, making it useful for feature selection. The mutual_info_regression() function in scikit-learn calculates the mutual information between each feature and the target variable for regression problems. Click here for library reference

SHAP: SHapley Additive exPlanations, a Responsible AI framework

Silhouette score: It is a metric used to evaluate the quality of cluster, how well each datapoint fits within its assigned cluster. This ranges from -1 to 1. Positive is better

SMILES: Simplified Molecular Input Line Entry System. Click here for the NIH reference to try a sample conversion

t-SNE: t-Distributed Stochastic Neighbourhood Embedding

WCSS: Within-cluster sum of squares (WCSS)

WCSS is also known as inertia, is a measure used in clustering algorithms like K-Means to evaluate the compactness of clusters. It represents the sum of the squared distances between each data point and its assigned cluster centroid

Elbow method: WCSS is used in conjunction with the Elbow method to determine the optimal number of clusters (k-value). Example below. The Elbow point communicates that six is an optimal cluster.

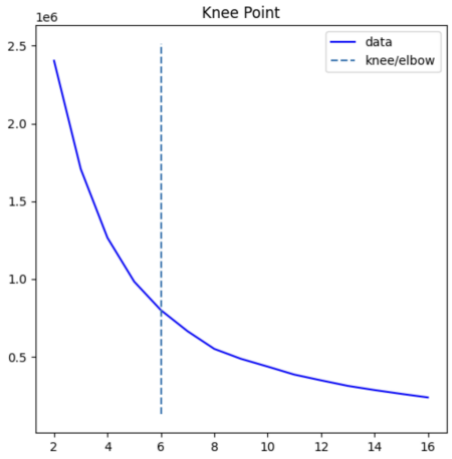


Figure 26: Plot identifying K value (Elbow or Knee point)